EAGLES User Manual

Contributing Authors: Kezia Manlove, Daniel Weiss, and Jennifer Sheldon

1.0 Overview

1.1 Project Objective and Intended Audience

This manual outlines a workflow and a set of software tools collectively known as the Ecosystem Assessment Geospatial Analysis & Landscape Evaluation System (EAGLES). EAGLES is designed to aid resource management decision making by providing support for species habitat planning efforts that integrate changing landscape conditions with demographic responses. Managers seeking to evaluate multiple development plan proposals can use this system to compare alternatives and scenarios, including changes in land-use practices, and explore their implications using hypothetical 'what-if' scenarios. For example, a manager could use this set of tools to investigate how coyotes currently use a portion of landscape, and how that use pattern might change when the landscape is altered (e.g., through fire, flood, or development). These tools are particularly relevant for legacy data on species of concern. To reach the widest possible audience, an ArcGIS environment was selected as the platform for these tools.

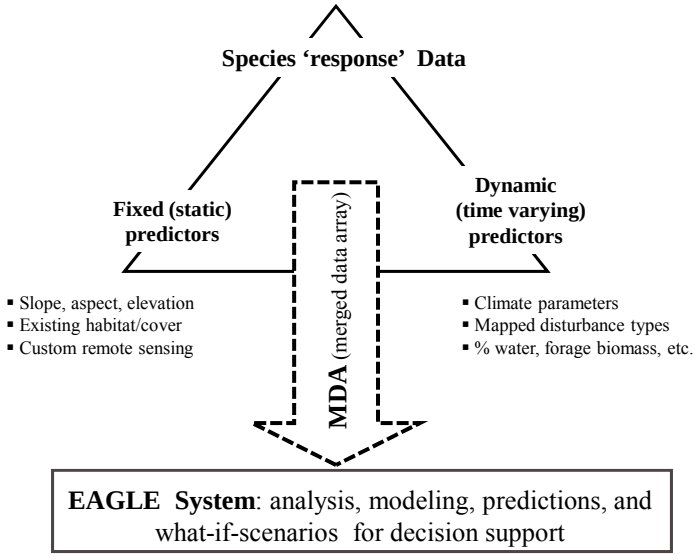


***Figure 1.1*:** Schematic representation of the process of matching (1) fixed and (2) temporally dynamic geospatial covariates with spatio-temporal response data from legacy data sets to create a merged data array (MDA) for analysis and modeling in EAGLES.

The workflow is designed for a user team with the following skill sets: GIS, basic knowledge of remote sensing data, access to a statistical consultant (for more complex decisions), and lead biologist/manager with expert-level species knowledge. These skills may be found in one person, but are more likely embodied in a group of people working in collaboration. This manual contains instructions useful to all members of the team and/or an individual user fulfilling all roles. Specific contents include an introduction to the ecological and statistical methodologies underlying the tools, an overview of the tools themselves and where they fall in the model-building workflow, and a worked example.

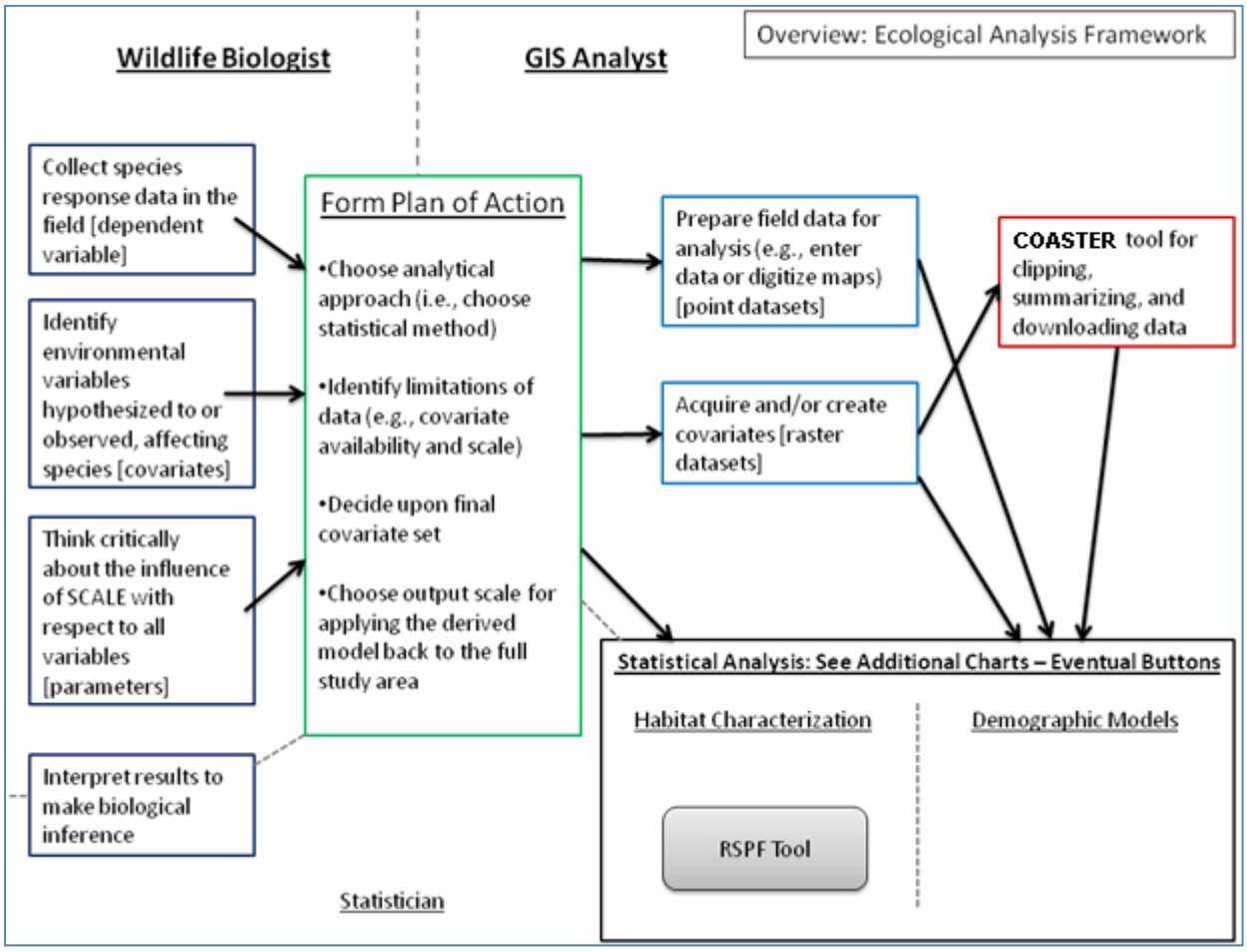


Figure 1.2: Example workflow for Resource selection (RSPF) analysis in EAGLES.

We expect that user/team will follow a work plan similar to the one outlined in Fig. 1.2. Our intent is to facilitate production of a model that is standardized, transparent, and defensible. We obtain each of these criteria as follows:

Standardization: By using hard-coded functions, we limit potential coding errors that might occur if each analysis was coded individually. We present a research framework and tool set that could be applied to many organisms and questions in many different systems.

Transparency: This manual contains relevant citations and methodological discussion, and the tools place heavy emphasis on visual display for the user, so that modeling assumptions can be clearly identified and verified.

Defensibility: In keeping with the Daubert paradigm for legally defensible science, we rely on well-documented methodologies (RSPF, etc.) with known sampling distributions, and thus quantifiable error rates and/or uncertainties. The workflow is designed to guide a user or team through steps via a series of dialogue boxes in the ArcGIS environment. While the tool itself embodies three primary functions (Data Input, Data Integration, and Analysis and Modeling; see Fig. 1.3), it is nested in a longer process of ecological investigation that begins with a set of management objectives and ends in ecological decision-making.

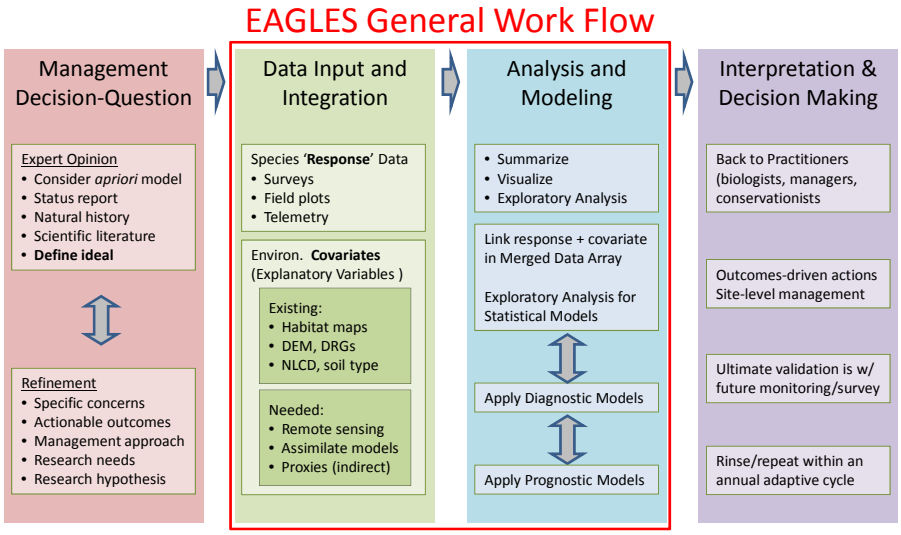


Figure 1.3: The EAGLES workflow schematic diagram. EAGLES is a workflow architecture that includes both tools (software based) and workflow to allow modeling of species legacy data sets to address management and conservation decision making. It is flexible and provides multiple workflow pathways based on the specifics of the species response data and management question(s). The general idea is to provide a systematic yet flexible architecture for integration of species data with geospatial covariates, most of which are derived from NASA data, data products, and ecosystem models that assimilate sensor data. As the degree of complexity in statistical analyses and remote sensing data increases, the need for a set of standardized techniques and common data protocols becomes more essential if we are to support repeatable, transparent methods for ecological modeling.

1.2 Development of a Narrative Model

This workflow is most effective when the user/team has a strong working knowledge of the organism of interest, including physiological drivers and potential thresholds, trophic roles including predators (hazards) and prey (resources), niche (competitive interactions) and habitat (geophysical) preferences, as well as parasites and disease (see Fig. 1.4). Our conceptual modeling process begins with a verbal description of important relationships (competitive, trophic, behavioral, etc.) between the organism of interest and its environment. This prior knowledge of the system enters the model through selection of a set of hypothetical drivers (covariates) to be considered for inclusion in the model. Here, we refer to the covariates and their relationship to the organism of interest as the narrative model. In order to aid the user in understanding an appropriate depth for the narrative model, we include descriptions of the narrative modeling in the tutorial.

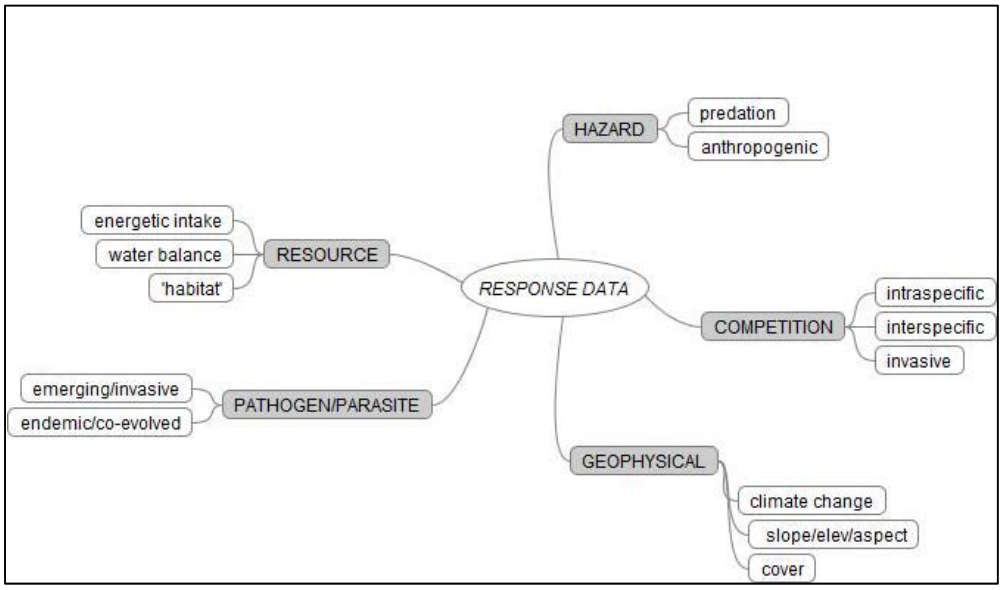


Figure 1.4: A mind map (Beel et al. 2009) visualization of various factors affecting variation in the focal species response or legacy data sets. Considering all possible risks and rewards based on expert opinion, research, and natural history helps avoid deficient models. It should also represent the ideal world of postulated mechanisms leading to testable hypotheses and management decisions. Covariates are then specified to represent these factors so that end-users can build a Merged Data Array (MDA) prior to data exploration tools, analysis, and modeling

1.3 Data Inputs

Data inputs can be classified into two broad groups: (1) species/population inputs (i.e., response data) such as GPS, radio collar data, survey and transect data, including flight data (2) geospatial covariate inputs, which may be derived from spaceborne sources such as MODIS and LandSAT,

airborne sources like LiDAR and NAIP, ground-based sources like meteorological base stations and distributed sensor networks, and modeled estimates like those from CASA, HYDRA and SRM. We use the term "response" here to refer to the species data used to fit the model. The model can then be extended to predict other species responses in addition to those actually observed and used for model fitting, in an effort to generate ecological forecasts.

In light of the plethora of new and emerging covariate inputs available, and the complexity associated with getting them into the ArcGIS environment, EAGLES provides the user team with two tools for data acquisition and formatting:

1) A wiki site that provides an index of existing geospatial covariates, as well as information on their contents and generation, located at http://geospatialdatawiki.wikidot.com/. Additionally, a partial list of frequently sought covariate layers and where to find them is included in Appendix 1 of this manual.

2) A tool to create climatic variables customized for the user's particular region on interest applicable for immediate use with the ArcGIS tool. This site can be accessed at http://www.coasterdata.net/.

1.4 Data Integration

The data integration portion of the analysis consists of accessing covariate layers and integrating them with the response data. In most cases, a Merged Data Array (MDA) is built and used for subsequent analysis. Functionality is provided within the ArcGIS-based tools to create the MDA, thereby relieving the user of several time consuming steps involved in preparing the data for direct export to a statistical program. Important considerations to keep in mind at this stage include:

1) Sampling approach and the distribution of response points
2) Spatial domain of analysis
3) Spatial scale of analysis
4) Modeled covariates

5) Availability space

Each of these topics is dealt with in detail in Section 4.

1.5 Data Exploration

Once the user/team has developed a narrative model and acquired covariates, the data exploration tools accessible as plot buttons in the first round of user dialogues in the ArcGIS RSPF tool provide a venue for preliminary data exploration and model fitting. These tools walk users through appropriate portions of the protocol for data exploration for ecologists proposed by Zuur *et al.* (2010) in order to better familiarize themselves with their datasets. This protocol consists primarily of graphical tools for identification of outlying data points, non-normal data distributions, and anomalies in data structure that should be considered in model selection and development.

1.6 Analysis and Modeling

EAGLES's statistical analyses occur in the statistical programming environment R. EAGLES currently has a Resource Selection Probability Function (RSPF) model, and a statistical model for intensity of use. More complex models that allow for mixed effects and spatial autocorrelation are under development. Due to the training required to effectively use R, the EAGLES workflow permits user interface in the more familiar and user-friendly ArcGIS environment. Expert users can also amend and interact with the underlying R code directly if so desired.

1.7 Model Assessment and Interpretation

Results from the preliminary data exploration and analysis both require a degree of statistical understanding to effectively build a model and interpret the results. A statistical consultation may be useful for many users at this stage, but users with even limited statistical training can assess results themselves by studying the examples provided in this manual and utilizing their knowledge of the species of interest.

2.0 Installation of ArcGIS Tools

The EAGLES tools are intended to assist users in acquiring data and fitting a Resource Selection Probability Function (RSPF) (Lele and Keim 2006; Lele 2009) in the open-source statistical computing environment R on a windows PC equipped with ArcGIS 9.X. The RSPF analysis tool utilizes statistical processing functionality contained in R scripts that are called directly from the ArcGIS interface. The intent of this tool is to provide users access to a powerful modeling framework without requiring extensive statistical programming knowledge.

2.1 Acquisition of Required Open Source Software

2.1a Installation Step 1: Download and install R

Download the latest version of R by navigating to http://cran.cnr.berkeley.edu/bin/windows/base/. When you follow this link, you will arrive at the site shown in Fig. 2.1. The version used in the worked example is R 2.10.1. Follow the *Download R 2.10.1 for Windows* link. The default installation is adequate for more users, and was used for the examples in this tutorial.

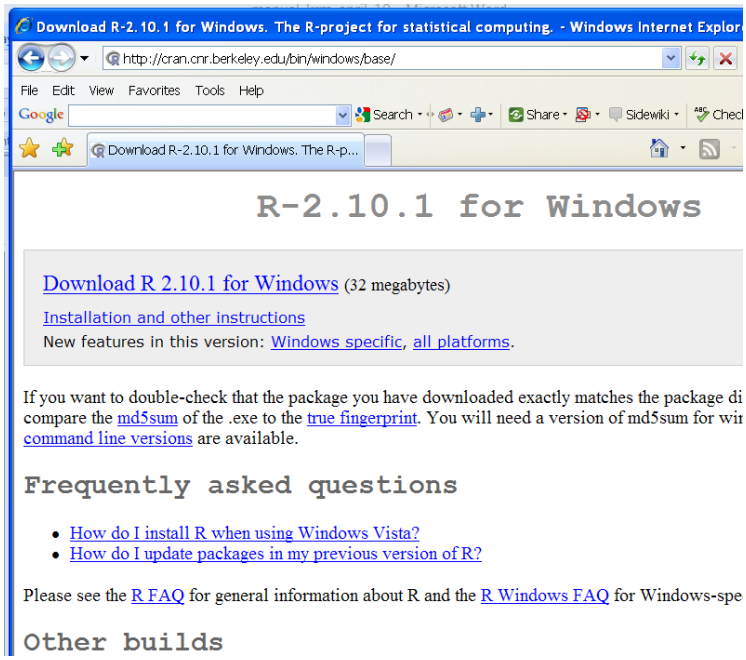


Figure 2.1: The webpage for downloading the R statistical software.

Unless fundamental changes are made to R, new versions of R should continue to work with the RSPF R scripts (this is not the case with the packages – see Installation Step 2)**.** The code has been tested on R 2.8.X, 2.9.X, 2.10.X, 2.11.X, and up to 2.12.1, however compatibility with future versions cannot be guaranteed.

2.1b Installation Step 2: load required R packages

The RSPF tool relies on functions housed in a variety of different R packages. Please note that several of these packages have undergone extensive reformatting since the inception of this project. While some users may have some or all of the necessary packages already installed on their machines, version updates make it necessary for all RSPF computations to be carried out through the set of packages provided. All necessary package files are contained within a designated folder in the zip file on the YERC website and are detailed in section 10.3.

To use the RSPF tool the contents of this library folder must be extracted and placed within the "library" folder of the R directory (e.g., c:\program files\R\R-2.10.1\library\).

2.1c Installation Step 3: Modify Windows Environmental Variables so R can be called by ArcGIS

After R is installed, Windows must be set to allow ArcGIS to start R. To do this, the user (with administrative access on the PC) must go to the computer's Control Panel and then to System Properties. Under the Advanced tab, click the button for Environmental Variables, as shown below. The Environmental Variables dialog window will open. In the variable list for System Variables, select path and hit the Edit button. Paste the path to the bin folder for R, located within R's program file (e.g., c:\program files\R\R-2.10.1\bin\), to the end of the Variable Value text line, separating the new path by adding a semicolon before pasting. This allows ArcGIS and R to communicate. Fig. 2.2 shows the windows that users will see on a Windows XP machine when setting the environmental variable.

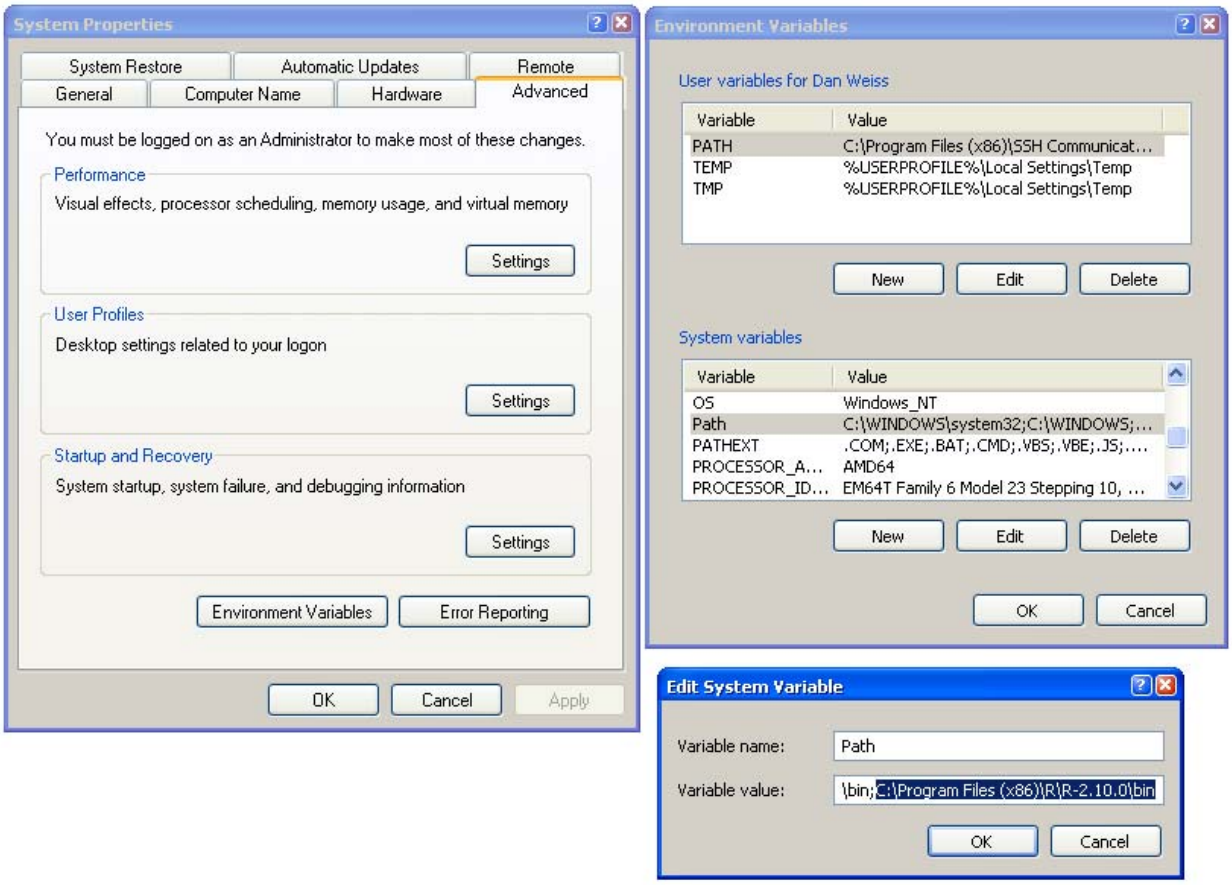


Figure 2.2: The windows opened when setting the environment path to allow ArcGIS to call R directly in Windows XP.

2.2 ArcGIS 9.X components of the RSPF tool

The RSPF tool may be started using two approaches. The first is to open the ArcGIS project file called RSPF.mxd. This file contains all the necessary code to use the tool as described in this manual. This approach is effective, but may necessitate copying the .mxd multiple times for various projects. Note that if this approach is taken, users are advised to clear the spatial information associated with the project prior to adding new data. To do this, go to view > Data Frame Properties > Coordinate System and click the "Clear" button.

Alternatively, users can set establish the RSPF tool as a clickable button that will be present every time ArcMap is started. Instructions for doing this are available in Appendix 4. Requisite files for both the .mxd and button installation are available for download from YERC.

3.0 Data Input

3.1 The Wiki Tool

The wiki tool (Fig. 3.1) allows a user/team to search for potential covariates using a variety of criteria (type of measurement, spatial scale, data source, etc.) and then links users to information on collection, acquisition and development of those data sources. This wiki can be updated by registered users, and is intended to function as a reference site for geospatial data (and particularly those data derived from remote sensing sources) that is of common interest to ecologists. To use the wiki, users can search by keywords or select from indexed lists of datasets described within the archive. Indexed lists can be accessed by pulling down the Main Pages menu found in the upper right hand corner of the screen and selecting Data Sets. Resulting lists appear as a set of links that can be clicked, thereby leading users to sub-lists and/or individual dataset pages.

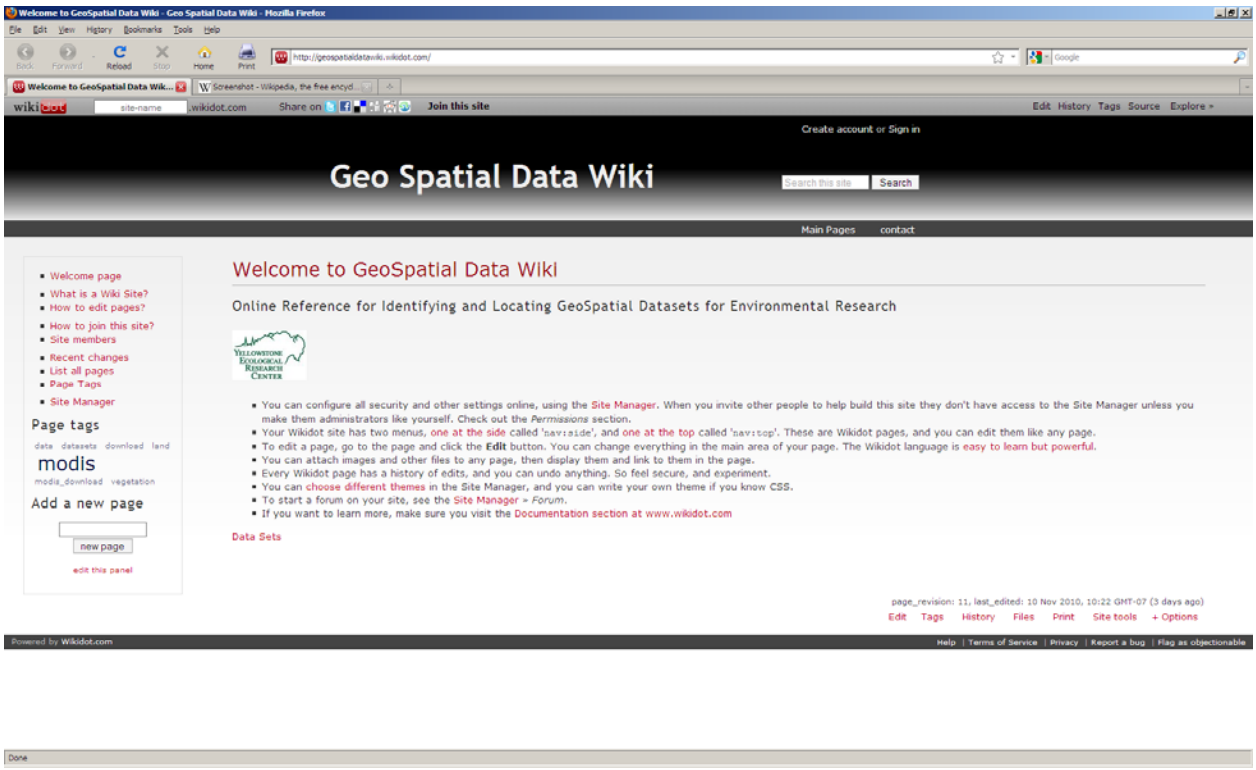


Figure 3.1: Screencapture of the GeoSpatial Data Wiki page.

The Geospatial Data Wiki can be accessed at http://geospatialdatawiki.wikidot.com/.

3.2 Customized Online Aggregation & Summarization Tool for Environmental Rasters (COASTER)

The COASTER system is a set of online tools designed to produce customized raster datasets for specific spatial domains. COASTER results can be used for data visualization and are amenable for use as input covariates in statistical models such as RSPF. The great strength of this approach lies in its ability to reduce massive and cumbersome datasets into manageable information that can be easily incorporated into an ArcGIS environment. The data currently available on the tool consist of gridded climate data for the Lower 48 United States, from 1980 through 2009, with an 8 km spatial resolution, and a daily temporal resolution.

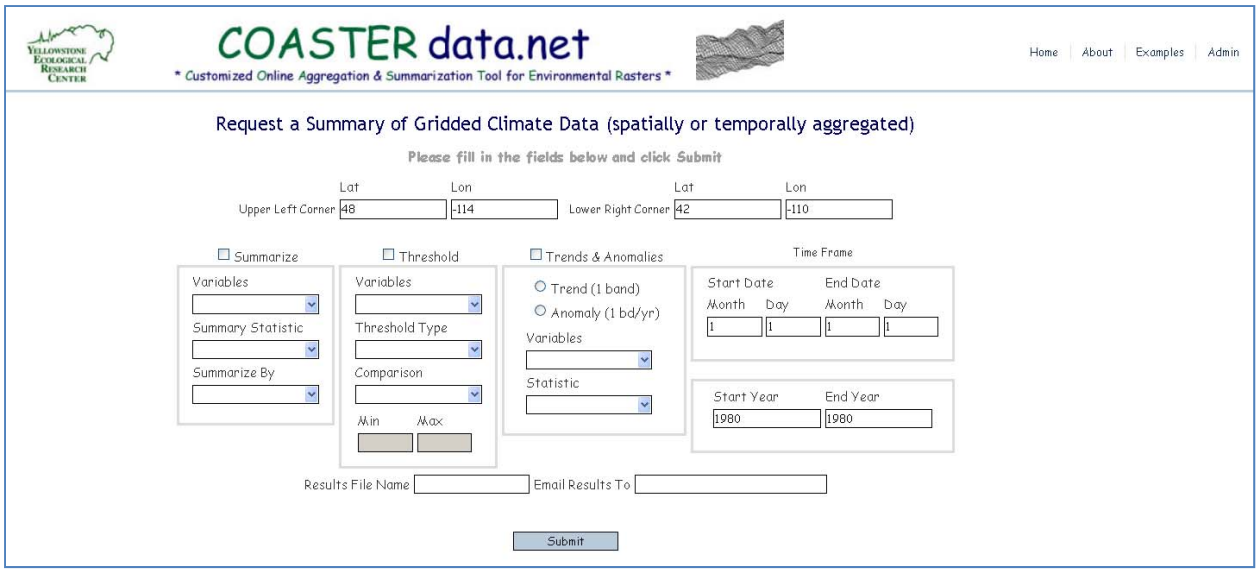


Figure 3.2: Screen-capture of the COASTER tool's information entry page.

COASTER is available at the following address: http://www.coasterdata.net/.

4.0 Data Integration

Before analysis in the RSPF tool can begin, the desired covariate and response layers must be projected and/or transformed into a common projection so that all raster cells are referenced within the same coordinate plane, and are thus properly aligned for data extraction. We also recommend that all datasets be resampled to the same pixel size (spatial resolution) before sampling takes place. Furthermore, the spatial extent of covariates must overlap the region of interest, as modeled output from EAGLES can only be created for areas possessing data for all covariates

4.1 Sampling approach and the distribution of response points

Sampling species responses to create a dataset capable of adequately addressing current and future research questions is a major undertaking that is exacerbated by the high cost and limited resources allocated for data collection. A tremendous amount of effort has gone into the study of sampling designs (for more information see Miller *et al.,* 2007 pg. 228 col. 2 bottom). Specific challenges noted in the geographic literature include:

1) Selecting the variable(s) to be collected that capture the necessary information using robust, repeatable, and defensible methodology.
2) Selecting the type of data (e.g., counts, presence/absence, or measurements of characteristics).
3) Selecting an underlying sampling approach (e.g., random, opportunistic, etc.) that does not violate the assumptions of the desired statistical methods.
4) Collecting a sufficient sample size.
5) Adequately accounting for the spatial distribution of sample points (i.e., points with high spatial proximity may be spatially autocorrelated and therefore of less informational value than points sufficiently far apart, whereas points too far apart introduce potential extrapolation error).
6) Planning long-term strategies (i.e., can equivalent data collection occur at multiple time periods to create a longitudinal dataset).

4.2 Spatial domain of analysis

Analysis may be conducted over a spatial extent that exceeds the area in which sample data were collected. While very powerful, this feature must be used cautiously as inference in unsampled areas, particularly areas dissimilar from any sampled points, should only be done with extreme caution as inference there is not well-supported by the statistical models. Spatial extents of potential interest (i.e., for management actions) could include politically defined units (e.g., hunting district or management areas) or geographically bounded regions (e.g., the Lamar Valley in Yellowstone National Park). Note that to apply a model to an entire landscape requires that all underlying covariate datasets have spatial extents that cover the entire area of interest.

4.3 Spatial scale of analysis

To produce the most interpretable results, all covariates entering the model should share a common spatial scale (i.e., spatial resolution). However, data layers are often collected or modeled at very different scales. Ideally all the covariates will have an identical spatial resolution (e.g., 30 meters) that is consistent with the spatial error term associated with the response data. This is seldom the case, however, and user teams will typically need to decide on a scale appropriate for analysis. In cases where covariate data must be rescaled, the user has two possible options, each of which has drawbacks. The options are:

(1) Scaling up – in this case the resolutions of the covariate datasets are reduced (i.e., multiple pixels are averaged to create coarser pixels) until they match the resolution of the coarsest dataset and/or the maximum spatial error of the response dataset. The advantage of this approach is that when the statistical model (e.g., the RSPF fit) is applied to the entire spatial domain, inference will never be made at a finer resolution than the datasets can allow. The drawback of this approach is the loss of detail in the covariate datasets that may have been costly to collect or acquire.

(2) Scaling down - in this case the resolutions of covariate datasets are increased to match the resolution of the finest dataset through the process of resampling. The advantage of this approach is that all data are preserved. The drawback is that when the RSPF model is applied to the full study area (e.g., the RSPF fit image is generated), inference is being

made at spatial scales unsupported by the input datasets (i.e., an ecological fallacy). Statistically this approach is much harder to defend than scaling-up, but there are occasions when it is more justified than others. For example, for a covariate that is at a coarser scale than desired is unlikely to vary significantly at a fine scale (e.g., temperature within a flat area such as a plain), scaling down may not compromise the analysis by adding excess noise through resampling the temperature covariate, and would allow other covariates to enter the model at their finer, more informative scales. However, users should be prepared for the tell-tale checkerboard effect (i.e., visible squares representing the grid cell boundaries of the original dataset) visible when the model is applied to the entire spatial domain.

4.4 Modeled Covariates

Many different modeled covariate layers are available for model inputs. Some of these are freely available while others are available for purchase (see Section 3.1 for use of the geospatial wiki for covariate identification and acquisition). Modeled covariates may include Digital Elevation Models (DEM) and their derivatives, modeled outputs from BioGeoChemical (BGC) models such as Biome-BGC, interpolated climate data produced using meteorological station data (e.g., PRISM, TOPS), and products made using remotely sensed imagery (i.e., images collected by airborne and spaceborne sensors and used to estimate values of ecological meaning). For example, Net Primary Production (NPP), mean winter precipitation, and forage biomass estimates are modeled covariates.

4.5 Availability Space

Availability space (i.e., places on the landscape where the sampled species could have been observed) is a necessary input for the RSPF tool. The user/team is responsible for determining an appropriate method for obtaining available points for their focal organism. Selection of available points remains an active research area. The tool provides three options for creating or importing availability points, which are detailed in Section 6.4.2. Generally we advocate that users create their own availability points to have greater control over their spatial distribution.

4.6 Merged Data Array (MDA)

The final product of the data integration phase is a merged data array (MDA), a table that can be read by a variety of different statistical programming environments. The MDA is created in ArcGIS by intersecting all response and availability points with each covariate raster dataset and extracting the covariate value for each use/availability point by spatial location and written as a .csv file. In EAGLES, the MDA is then passed to the statistical programming environment R for analysis, though the user/team could read it into any statistical programming environment they chose. While the R processing will be automated for ArcGIS users, the underlying R code is available for user inspection and modification.

5.0 Data Exploration

We provide a set of data exploration tools, available after the MDA is first sent to R (i.e., on the graphs available in the window that pops up after users click the first "Submit" button). The data exploration portion of the analysis is intended to help the user team familiarize themselves with the data. Specifically, we advocate the use of a portion of the data exploration presented protocol by Zuur *et al.* (2010) that is as follows:

1) Identify potential outliers in all covariates and the response (through use of boxplots and histograms)
2) Look for collinearity in the covariates using the pairs plot
3) Look for relationships between the covariates and the response
4) Examine independence assumptions in the response using semivariograms
5) Examine spatial distribution of the covariate values to spatially identify unusual regions

Since we anticipate that our audience will often be using a count or binary response, we advocate a post-model-fitting assessment of normality using a normal quantile-quantile plot (in development).

5.1 Pairs Plot

EAGLES produces a standard pairs plot (Fig. 5.1) that contains a great deal of information about univariate and bivariate distributions within the dataset. On the main diagonal of the plot matrix are histograms of each covariate, where the user can look for outlying points and multimodality (that is, multiple peaks in the distribution). To spot outliers, look for histograms that have long tails. The upper triangle of the plot matrix contains pairwise scatterplots of all covariates. Use this plot to identify potentially collinear variables. Collinear variables are variables that have strong relationships with one another, identifiable by the points in the scatterplot all falling along a line. When two collinear covariates are both included in a model, the model fitting algorithms cannot identify which variable actually drives the response, which may result in the misallocation of influence to one covariate or the other. When collinear covariates are of interest, we encourage the biologist to make a decision based on prior knowledge of the system about which covariate is most logical for inclusion in the model. Here we can see that several

covariates, for example June NPP and May NPP, are highly collinear, which might suggest that only one of them should be used in our final model.

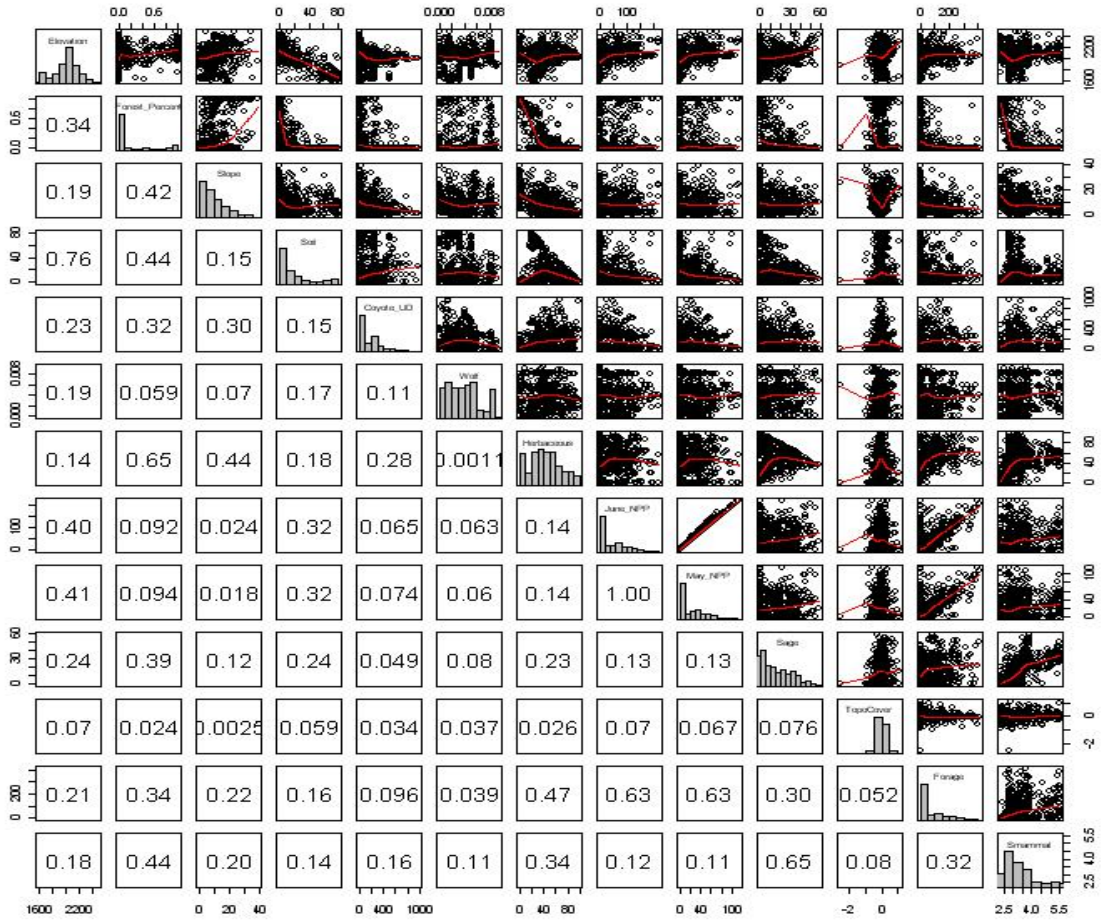


Figure 5.1: Standard pairsplot.

5.2 Conditional Histograms

Stacked histograms can be used to compare the distribution of a covariate at used, available and universal (that is, entire study domain) scales. In Fig. 5.2, which shows stacked histograms for elevation, we see that the full spatial domain extends quite a lot higher than either points that were used or points that were deemed available. In this case, the region with the highest elevation generally resides outside of the area that is modeled, thus inference to very high elevations is beyond this model's scope.

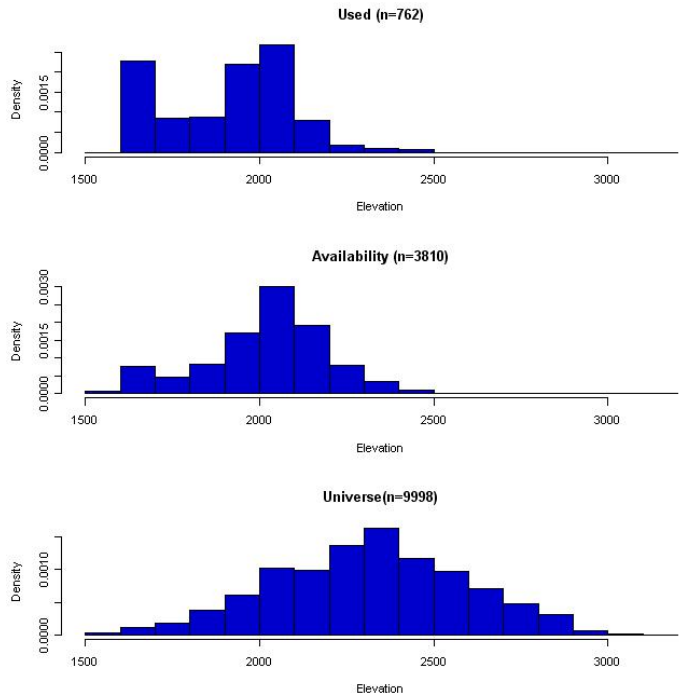


Figure 5.2: EAGLE conditional histograms.

5.3 Semivariograms for Assessment of Spatial Scale

Semivariograms are useful for showing the spatial scale at which spatial autocorrelation is present (or not) within an environmental covariate. The presence of spatial autocorrelation is normal in environmental datasets, but must be considered when interpreting results. Figure 5.3 shows a semivariogram in which autocorrelation ceases to be an issue for this variable at ~8000 meters (i.e., the sill, or fairly constant horizontal section of the semivariogram, begins at about this distance). Within univariate semivariograms autocorrelation is particularly problematic when there is no obvious sill (e.g. a linear decrease in autocorrelation with increasing distance).

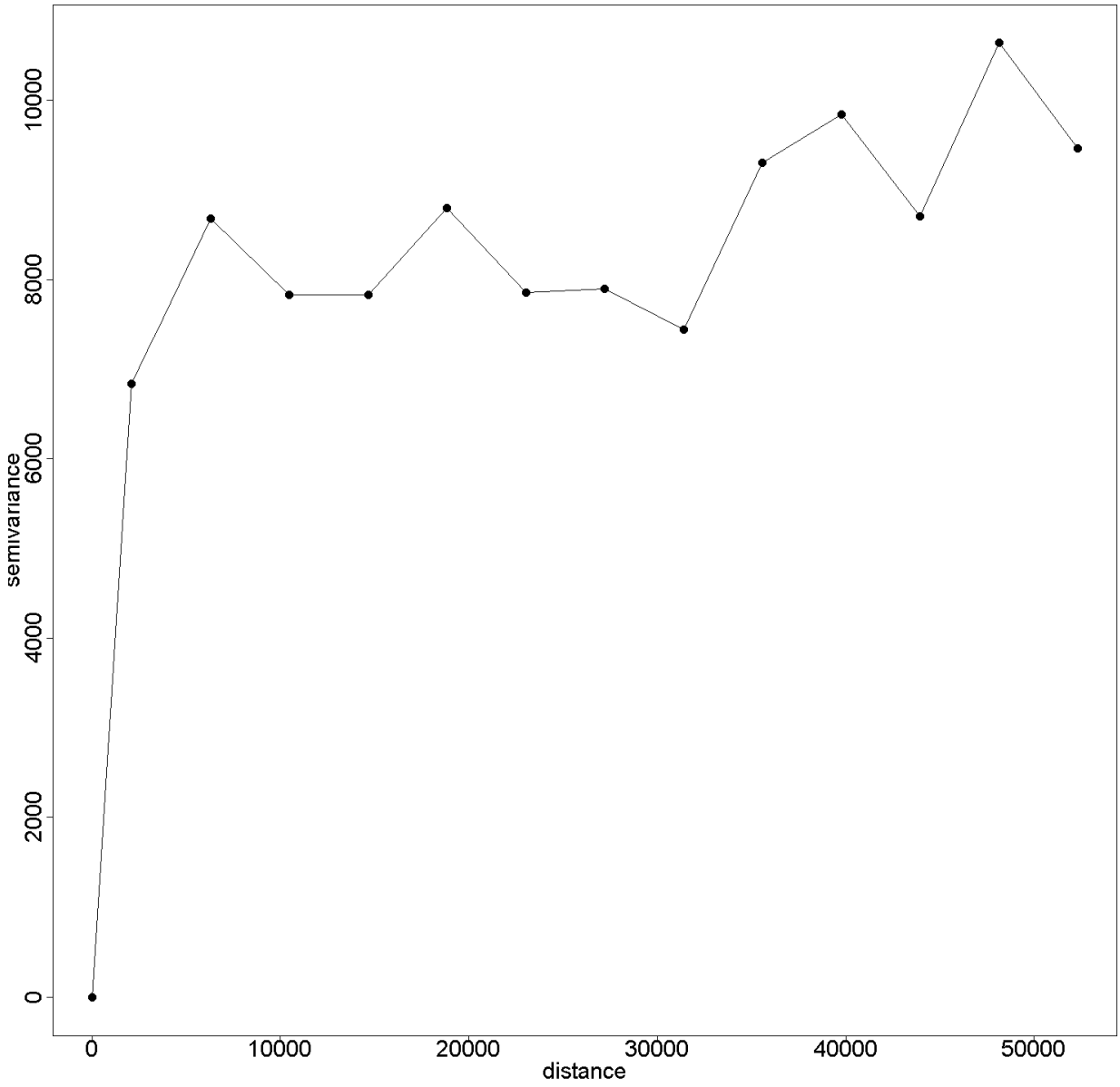


Figure 5.3: EAGLE univariate semivariogram.

5.4 Spatial Distribution

The spatial distribution graphic is useful for identifying the location of points that are outliers in the covariates. Knowledge about this spatial organization facilitates more informed, landscape specific interpretations of results based on expert knowledge. Because the graphic, as made in R (Figure 5.4), is somewhat rudimentary the bin values for each variable are included within the

table “rspf_used_points_with_bins.csv” and can be used to add the associated bin value for each use point within ArcGIS (i.e., join the .csv file to the use point shapefile attribute table).

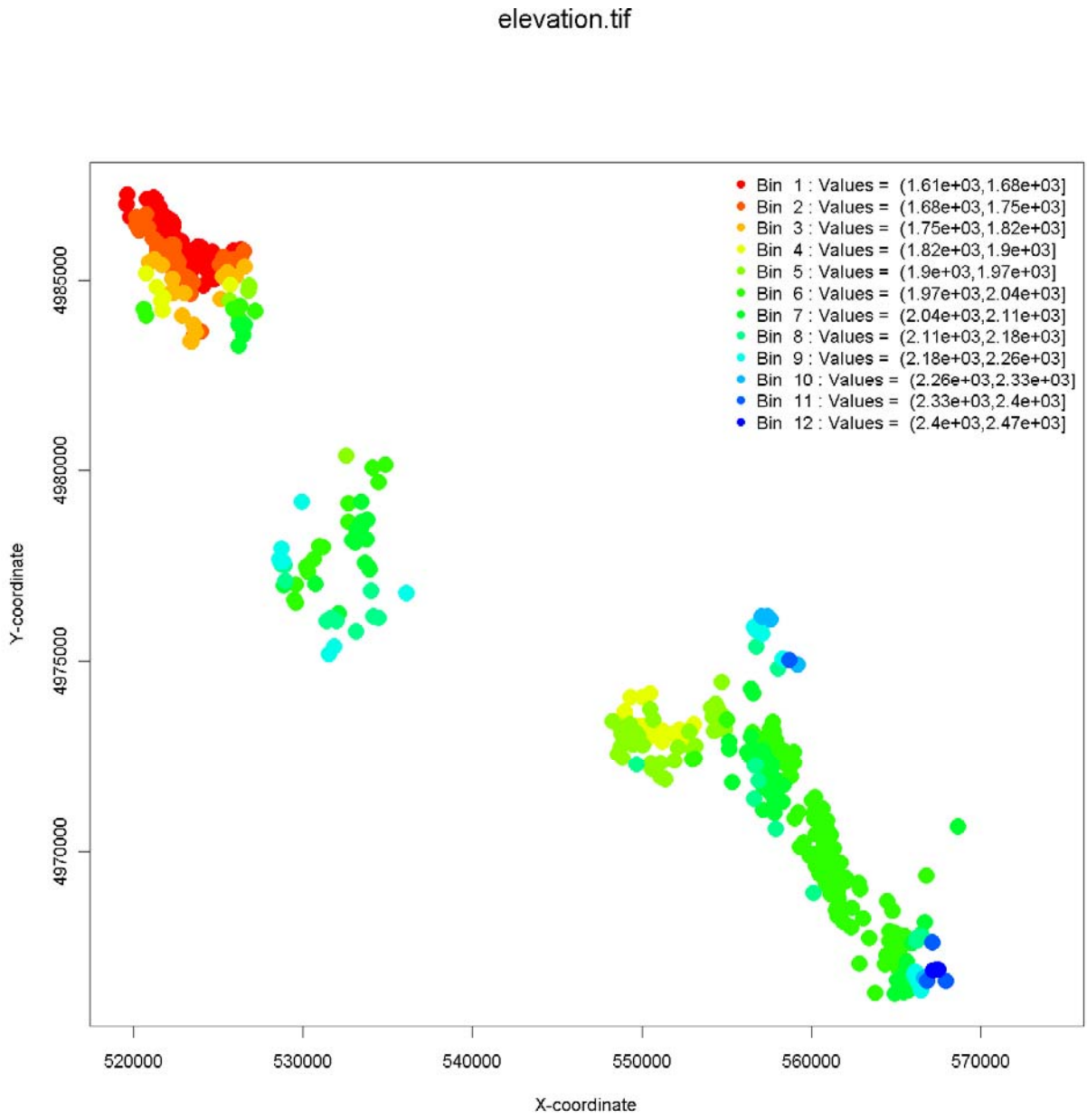


Figure 5.4: EAGLE spatial distribution for single covariates.

6.0 Resource Selection Probability Function (RSPF) Tool

The RSPF tool fits resource selection probability functions, a special class of species distribution models for use/available data and a set of desired covariates, directly from ArcGIS. Species distribution models (SDMs) are commonly used in ecological studies to characterize the relationship between the regions utilized by a species and the habitat features that characterize those regions. One specific manifestation of species distribution modeling is the Resource Selection Probability Function (RSPF) (Lele and Keim, 2006; Lele 2009). RSPF is a model that estimates the relationship between habitat use and attributes of important covariates through a model akin to standard binomial regression models (logistic, cumulative log-log, etc.). While it is our intent to make a variety of other species distribution models available, we developed the initial modeling code for the logistic RSPF due to the ubiquity and transparency of its underlying logistic regression model.

6.1 Introduction to RSPF

RSPF modeling is an extension of Resource Selection Modeling that relies on resampling theory to resolve problems associated with obtaining truly "Unused" points. RSPF methodology does not imply a particular link function; rather, it adjusts model standard errors so that they accommodate Use-Available, as opposed to Use-Nonuse sampling designs, by considering used points to be draws from the following weighted distribution,

$$f^U(\boldsymbol{x};\beta) = \frac{\pi(\boldsymbol{x};\beta) f^A(\boldsymbol{x})}{\int \pi(\boldsymbol{x};\beta) f^A(\boldsymbol{x}) d\boldsymbol{x}}$$

where $f^A(\boldsymbol{x})$ is the distribution of covariates for the available population, $\pi(\boldsymbol{x};\beta)$ is the resource selection probability function, and $\int \pi(\boldsymbol{x};\beta) f^A(\boldsymbol{x}) d\boldsymbol{x}$ is the expected probability of use. RSPF estimation allows us to estimate β in $\pi(\boldsymbol{x};\beta)$ based simply on $f^U(\boldsymbol{x};\beta)$, the distribution of covariates in the used population (Lele and Keim, 2006). We developed all models on RSPF functions with logistic links, however a cumulative log-log link (for the RSPF analog of a proportional hazards model) is available as well.

6.2 RSPF Tool Description: Code, data, and output file storage protocol

Code:

After creation of the MDA, the ArcGIS script calls two R scripts (presently called "RSPF_script_1.r and RSPF_script_2.r). The location of the R scripts is set by users when they run the RSPF tool. Since the RSPF tool calls the two R scripts by name, *please do not rename the scripts.*

Data Files:

The datasets used as inputs by the RSPF tool are made available to the RSPF tool by adding them to the ArcGIS project. The user should be aware that due to file reading structures within R, data file names must begin with a letter. *Do not begin a file name with a numeric character.*

Output Files:

The output files produced by the RSPF tool will be placed in a user-defined output folder. Within this folder, sub-folders will be created named "RunX", where X is a count that will increase by one for each user run. For example, if the user runs the RSPF tool three times, the output folder will contain sub-folders named (Run, Run1, and Run2). Contents of the RunX folders are discussed in Sections 6.3 and 6.5. *Due to the structure of the R environment, users must avoid spaces in file names as well as numerals in the first value of the filename e.g. "1animal".* Inside the RunX folder, the user finds the following four subfolders:

1) Parameters: contains the parameter files written by ArcGIS and read by R
2) Covariate Graphs: contains jpegs of all images displayed in ArcGIS, as well as several additional diagnostic plots
3) Results: contains model summaries, including coefficient estimates and statistics associated with model fit
4) Tables: contains the used, available, and universe csvs written by ArcGIS, as well as the RSPF model matrix, which is the MDA.

6.3 RSPF Tool Description: RSPF tool data flow and processing overview

(1) The RSPF tool operates as a GIS-based Graphical User Interface (GUI), collects user-defined information (e.g., input file names and output file destinations), creates a Merged Data Array (MDA) by extracting values from each raster dataset for all use and availability points, and generates a parameters file (visible to the user in Parameters sub-folder of the RunX folder, in the file RSPF_params_aks.txt) allowing these arguments to be passed to the first R script.

To accommodate the requirements for extracting the raster values for each point, the RSPF tool may resample raster files according to the user-specified scale of analysis. As a result, a (potentially) modified version of each input file may be placed in the RunX folder. Since the input files can be quite large, this procedure has the potential to take up large amounts of disk space.

(2) R is called directly by ArcGIS and the first R script is executed using the MDA created in step 1 and arguments specified in the parameter file. The first R script derives an empirical univariate RSPF for each covariate, diagnostic graphs for data exploration, and graphical and tabular output for each selected covariate, and enables the user to select covariates contributing to the final, composite RSPF fit. At this point, we reiterate that only one in a pair of collinear covariates should be used in the final fit.

(3) The output files generated by the first R script are sent back to the ArcGIS tool to allow for additional user specification of the RSPF model (e.g., the user will select which covariates to include in the final model, as well as the desired link function). The RSPF tool now generates a new parameter file (overwriting the existing parameter file in the process) that is passed to the second R script. The new parameter file is very similar to the first, but also contains the new user-selected arguments.

(4) R is called from ArcGIS again and the second R script is executed using the arguments specified in the updated parameter file. The second R script produces an output RSPF fit based on the user-provided response and availability points (discussed in Section 4.4), and writes the equation used for fitting in the RSPF_equation.txt file.

(5) ArcGIS reads the RSPF equation from the RSPF_equation.txt file and applies the model to the entire raster dataset from which the MDA was derived. The final result is a raster layer depicting the RSPF model fit for the landscape. Details on how to interpret this dataset are provided in Section 7.

6.4.0 RSPF Inputs Overview

All input files used by the RSPF tool must be spatial datasets formatted as (1) vector shapefiles for the use (i.e., response) and availability datasets or (2) raster datasets for the environmental covariate layers. The RSPF tool was tested primarily using .tif images, and use with other raster data formats may behave unexpectedly. All input data files should be in the same datum (e.g., WGS-1984) and projection (if the data are projected, the model will also work using data in geographic coordinates).

6.4.1 RSPF Inputs: Response Data (i.e., Use Points)

The Use point shapefile contains points that represent known location of the species of interest (i.e., observations, telemetry locations, or GPS collar locations) from the sampling period.

6.4.2 RSPF Inputs: Availability Points

Availability points are used to define potential habitat. In other words, these are points where the species of interest *may* occur within the study area. In practice, however, the expertise of the researcher is often used to define a logical available space based on their understanding of the species of interest. See Forester *et al.* (2009) for a discussion of availability.

The RSPF tool provides users with three different options for identifying availability points, each of which corresponds to a different level of user control and thus a different combination of bias and variance in the model error structure. The RSPF tool availability point options are:

(1) The point buffer option, which generates random availability points located within a buffered region around each response point. In this case, the user must define both the buffer size and the number of points per buffer region.

(2) Random selection of a user-defined number of availability points from a region of available space (i.e., a polygon shapefile) regardless of the observed distribution of response points within the region. For example, the tool can pick five times the number of use points uniformly from an entire available space. This method leads to a uniform sampling intensity over the entire available region.

(3) (Preferable) The user defines availability points within a custom-made shapefile and enters these points directly into the model. The benefit of this approach is control, since the user can purposefully exclude points from areas that are not truly available (e.g., water bodies for terrestrial species).

We note that the appropriate number of availability points and their spatial distribution remain somewhat nebulous issues. A rule of thumb for the number of availability points is to use five times the number of response points. The spatial distribution of availability points is typically random within a defined availability space. A more complex issue is the distribution of points within disconnected areas of available habitat (e.g., three "patches" of habitat with varying numbers of response points within each). The preferred approach in such a case is to distribute availability points within each area in proportion to the number of response points. This approach, however, requires slightly more GIS acumen to produce than a simple random distribution.

6.4.3 RSPF Inputs: Environmental Covariates

There are three primary concerns when selecting and/or preparing covariate datasets for use with the RSPF tool:

(1) The raster datasets should underlie all the use and availability points. If this is not the case, the Merged Data Array (MDA) will contain inappropriate zero values that will be automatically removed by R, so as not to impact the validity of the statistical output, but no feedback is provided to the users indicating that they included invalid points. Alternatively, users can select an option to omit and points outside the region of interest prior to creating the MDA .

(2) The covariate datasets should have the same spatial resolution prior to analysis. Ideally the user will generate these manually to be aware of the important decision points made when resampling. The RSPF tool, however, has the ability to resample raster layers to an identical, user selected resolution. See Section 4.5 for a discussion of whether to scale up or down.

(3) The raster datasets should have the same spatial extent, as the RSPF model will be applied to the entire user-defined ROI. This is an important consideration because if covariates have mismatched extents, some areas in the resulting RSPF_fit raster will be generated without all the necessary covariates.

6.5.0 RSPF Fit: Fitting the Univariate RSPFs In R

In summary, the first script file fits a univariate Resource Selection Probability Function (RSPF) for each of the submitted covariates. These functions are of the form

$$f(\mu(y)) = \beta_0 + \beta_1 x,$$
$$f(\mu(y)) = \beta_0 + \beta_1 x + \beta_2 x^2, \text{ and}$$
$$f(\mu(y)) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

where $f(\mu(y))$ is a particular link function relating the probability of use to given levels of the covariate, denoted here as "x". This approach to model fitting allows users to select both an appropriate order for each covariate and an appropriate link function for their final model through examination of these first round models during the second user dialog. Note that the same link function must be used for all covariates, so we suggest that the user make two passes through the univariate RSPF plots. In the first pass, assess which link appears to perform best across all covariates, and select a link function for use in the final model. Then, once the link function has been selected, make a second pass through the plots, and identify the best order of fit for each desired covariate, considering only the fits based on the chosen link. In general, we give preferential treatment to the logistic link due to its interpretability and ubiquity.

In R, the models are fit using the Nelder-Mead (N-M) algorithm, a commonly-used simplex method that searches for optimal parameter estimates by finding a minimum in the multidimensional parameter space. In this case, initial values for the N-M algorithm are parameter estimates generated by the expectation-maximization (EM) algorithm used in fitting a generalized linear model (that is, the starting points for establishing optimal values for the weighted distributions are fits from an unweighted generalized linear model using the same link). The N-M fitting method relies on unimodality, and may fail in situations with multiple local minima. Sometimes convergence of the N-M algorithm is very slow. If convergence has not been achieved in 5000 iterations of the N-M algorithm, the RSPF function may be fit via Simulated Annealing (SANN), an alternative optimization algorithm that works well on rough surfaces. Users are informed when simulated annealing is employed in R, but should not be concerned by its use. For an introduction to link functions and covariate selection in resource selection models, see Manly *et al.*, 2002. For the original work on the Nelder-Mead algorithm, see Nelder and Mead 1965.

6.5.1 RSPF Fit: Determining Covariate Presence and Order for the Full Model

After obtaining some knowledge of how each covariate relates individually to resource selection, the user may wish to construct one or more multi-covariate models, composed of a combination of covariates each fit at some particular order. The RSPF tool incorporates several measures for assessing model fit and performing model comparison to facilitate multiple regression model fitting and selection.

1) AIC (Akaike's Information Criterion), a measure that represents a compromise between the likelihood of a particular parameter set and the number of parameters used to fit the model. Low values of AIC are preferable, though models are typically taken to be similar in functionality if their AICs are within two units of one another (see Burnham and Anderson).

2) AUC (Area Under the Curve). AUC is derived from the Receiver Operating Characteristic (ROC) curve associated with a model. The ROC curve shows the trade-off in the model between specificity and sensitivity (that is, it shows how often

the model predicts false positives and false negatives). In general, higher values of AUC correspond to models that exhibit more desirable properties with respect to both specificity and sensitivity. See Section 7.1 for additional information on ROC curves and AUC.

3) Goodness-of-fit: Various goodness-of-fit measures have been proposed for binary and binomial response data. Here, we use them in a Use-Availability setting, which is not exactly binomial, but the measures should work fairly well nonetheless. While none of these measures are without their caveats, a commonly used statistic is the Hosmer-Lemeshow statistic, which essentially bins the data over values of the covariate, and then uses a Chi-square test to compare observed counts in a bin to counts expected in that bin under the model. Since the null hypothesis for the Hosmer-Lemeshow test is that the model fits well, low p-values correspond to *lack of fit* in the model. An alternative method is the Kolmogorov-Smirnoff (K-S) goodness of fit test, which is a very general test of the difference between the used and available distributions of that covariate.

In addition to these statistical measures, the user must rely on knowledge of the biological system at hand, as well as the information contained in the curves, to select the order at which each covariate should be fit. For example, a covariate whose optimal values for a given organism are at the middle of the covariate's value range might be a candidate for a quadratic (second order) term, whereas a covariate whose optimal values for the organism are at the low end of the covariate's range, and whose increasing presence corresponds to steadily declining desirability might be a good candidate for a linear (first order) fit.

6.5.2 RSPF FIT: Fitting the Full RSPF in R

The second R script uses the N-M algorithm (or simulated annealing if appropriate – see above) to fit a RSPF function for the particular covariates and covariate orders specified in the second user dialog. Fitting here works the same as it did in the first R script, but only one model is fit. This model is of the form

$$f(\mu(y)) = \beta_0 + \beta_{11}x_1 + \cdots + \beta_{21}x_2 + \cdots + \beta_{k1}x_k + \cdots$$

where $f(\mu(y))$ represents the link function for the mean that the user selected in the second user dialog, β_0 is an intercept term, and $\beta_{j1}x_j + \cdots$ represent all terms related to the jth selected covariate (this could potentially include as many as three terms, of the form $\beta_{j1}x_j + \beta_{j2}x_j^2 + \beta_{j3}x_j^3$). Parameter estimates and fit statistics associated with this model are available to the user in the rspf_fit_summary.txt file located in the Results subfolder of RunX folder.

6.5.2a: Standardization

The user should be aware that all quantitative predictor variables are standardized prior to fitting. While standardization is a transformation procedure that does not affect the model fit or predictions, it does facilitate model interpretability (Gelman and Hill, 2007, pg. 56). Examples of appropriate interpretation of standardized coefficient estimates from the logistic RSPF are included in the worked example.

6.5.2b: Interaction

While the EAGLE tools do not generate interaction terms internally in R, the user can readily generate interaction layers in ArcGIS and pass them to the R models. We suggest the following guidelines when working with interaction terms:

1) Consider the use of an interaction term for main effects that have large values.
2) When building interaction layers, note that the EAGLE tools rely on standardization prior to generation of higher-order terms. To be consistent, the user should first standardize the two layers he or she wishes to include in the interaction (by subtracting the layer mean and dividing by the layer standard deviation) and then multiply the two layers together to form the product layer.
3) Once the individual variables are standardized, an interaction layer can be created by multiplying the raster layers together using ArcGIS functionality such as the Raster Calculator.

7.0 RSPF Model Assessment and Interpretation

In order for a model to be scientifically defensible, it should meet two criteria:

1) It should be the best model of a suite of possible models
2) It should provide an adequate fit of the data

The EAGLES tools provide the user team with mechanisms for addressing both of these criteria. To assess criterion 1, we provide a model AIC value for the final RSPF model fit. We suggest that the user team generate a set of candidate models, fit each of the models in a series of runs of the EAGLES tool, and compare the resulting models in terms of their AIC values. The best model is the one with the lowest AIC. In order for a model to be deemed universally the best model, it should be two AIC points lower than the next-best model.

To examine criterion 2, we offer the user a measure ("area under the curve") and two plots. The plots are a Receiver Operating Characteristic (ROC) plot to examine the model's ability to correctly classify used and available points in the original dataset, and a semivariogram of deviance residuals, to assess whether the response exhibits spatial autocorrelation beyond that which can be explained by spatial clustering of the covariates.

7.1 Model Assessment: Receiver Operating Curve, Semivariogram for Spatial Autocorrelation, Goodness-of-Fit Statistics, and Model Coefficients

The Receiver Operating Characteristic (ROC) (Fig. 7.1) is a depiction of the probability that the model ranks points that were actually used as more likely for use than points that were not actually used (i.e., it is a measure of the model's ability to identify used points as used points and available points as available points). Higher probabilities indicate better models. The ROC curve is often summarized in terms of the Area Under the Curve (AUC, reported in the ROC legend). Higher values of AUC correspond to higher probabilities that the model classifies appropriately. If the model's classification is no improvement on random classification, then the ROC curve should sit at a line of slope 1 (i.e., the grey line in the background of the plot).

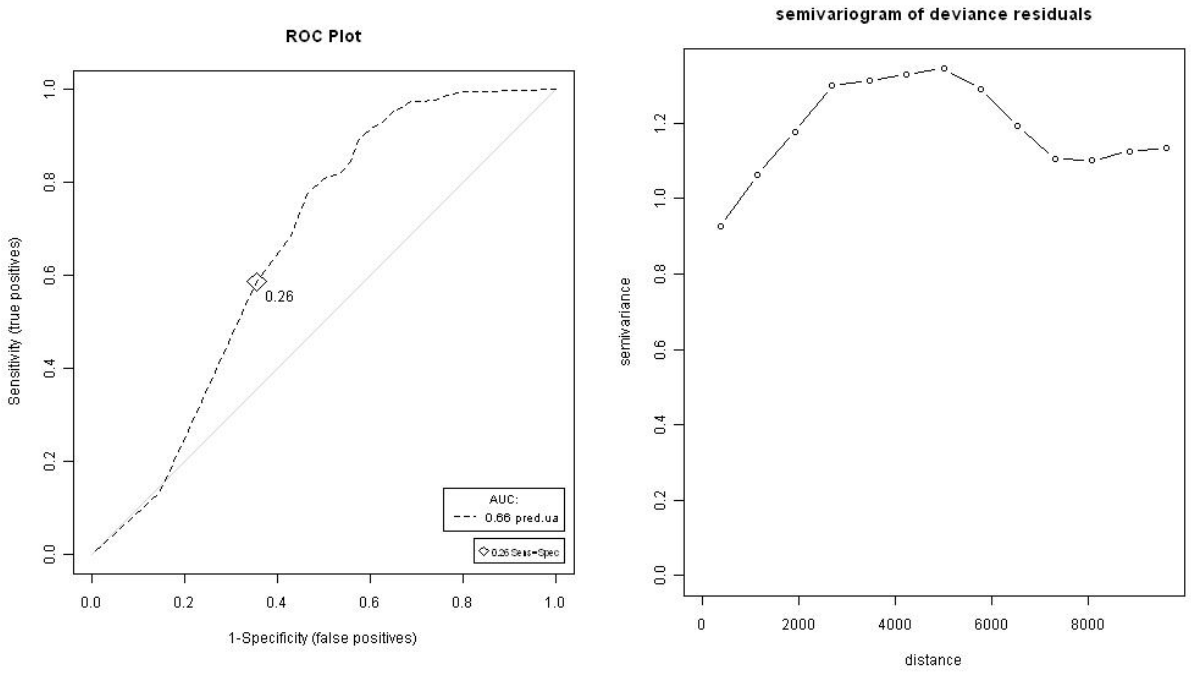


Figure 7.1: RSPF's ROC plot and semivariogram outputs.

The semivariogram is used to help determine whether lack of independence due to spatial autocorrelation is relevant in the setting of interest. This depiction of spatial autocorrelation relies on assumptions of stationarity (spatial relationships are the same over the entire spatial domain), ergodicity, and isotropy (spatial relationships are the same in all directions) for the underlying spatial process (see Zuur *et al.,* 2007 pg. 344). It is a plot of the variation between two distances as a function of the distance between two points. Ideally, we want this plot to be a horizontal line, which is indicative of similar variance between points regardless of the distance between them. Lower values of semivariance for lower distances indicate relatedness between spatially proximal points, which suggests a violation of the independence assumption in the model fit. Such violations necessitate the use of a more complex model, and if left unaccounted for, they may result in inflated Type I error rates (that is, they may increase the chance that users identify covariates as significant when in fact they are not).

Goodness-of-fit statistics provide a formal measure of model fit. These statistics are located in the RSPF_fit_summary file produced and stored in the Results subfolder of the RunX folder after

the second RSPF script has run. We provide a Hosmer-Lemeshow test statistic often used for assessing fit of binary regression models. The hypotheses for this test are as follows:

H_0: The model fits adequately

H_A: The model does not provide an adequate fit of the data

Small p-values for the Hosmer-Lemeshow test indicate some lack-of-fit to the model, however, this test is somewhat conservative. We provide this test statistic simply due to its historical impact in binary regression settings, and encourage users to rely on the ROC curves and especially AIC values as better indicators of the performance of one model relative to the suite of models of interest.

A brief assessment of model coefficients is prudent at this point. We anticipate that most users will use RSPF models with logit links, where the relationship between changes in the covariate and the mean response probability are exponential. As such, very large coefficient estimates should be regarded with a grain of salt, since they indicate massive changes in response probability with changing values of the covariate. Additionally, we recommend that the users check the variance inflation factors (VIFs) reported in the coefficients table. VIFs in excess of ten are indicative of problems with model fit, often related to multicollinearity among selected model covariates (see Pronghorn worked example, Section 9). If high VIF values are present, we suggest that the user(s) revisit the pairs plot in the first round of user dialogues in an effort to identify potentially collinear variable pairs. If such pairs can be identified, we recommend the exclusion of one of the paired covariates from the final model fit.

The final (and perhaps most important and intuitive) tool for model assessment is an examination of the fitted RSPF surface in ArcGIS (Figure 7.2). We recommend that users take a critical look at the fitted surface, and apply their knowledge of the ecology of the focal organism to assess whether the surface returned by the model makes sense. It is our experience that examination of the fitted surface can be useful in identifying important and overlooked covariates. If the fitted surface makes ecological sense, the ROC values are acceptable, the AIC score is the best (or

among the best) in the suite of plausible models, and the semivariogram does not display major departures from spatial independence, the model should be regarded as acceptable.

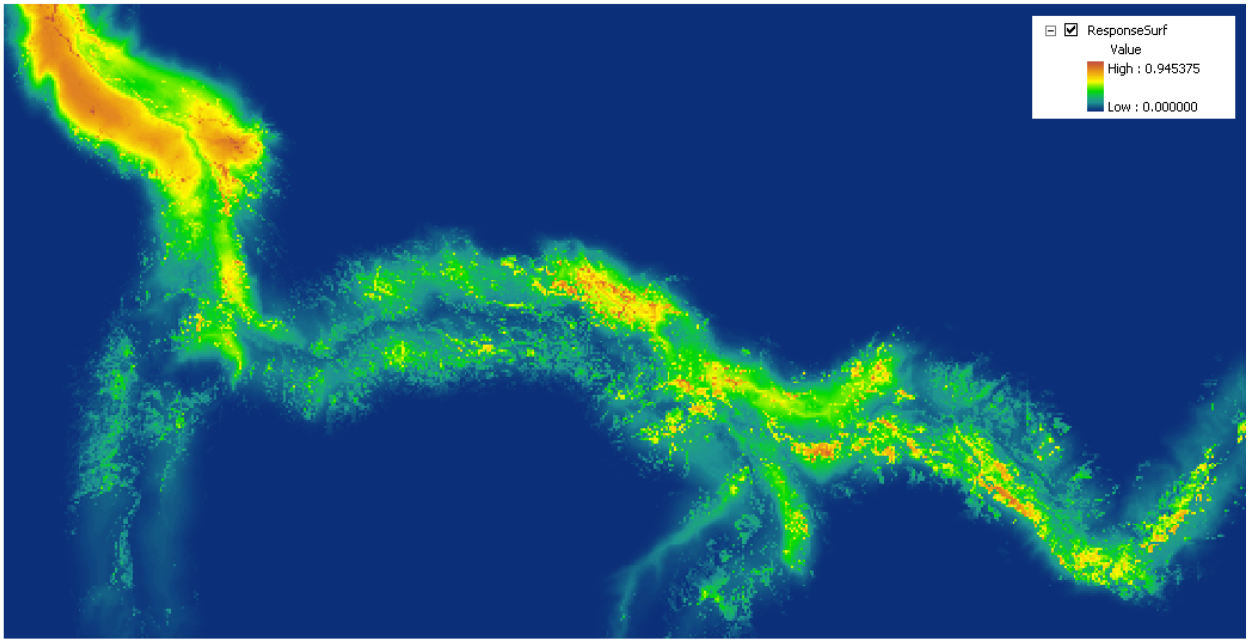


Figure 7.2: An example RSPF_Fit surface showing the probability that each raster cell will be selected by a species. Probability values range from 0 to 1 (i.e., zero percent chance of being selected to one hundred percent of the cell being selected according to the model).

8.0 Ecological Forecasting Through RSPF

The EAGLES tool provides functionality that allows users to apply RSPF models fit using observed data to potential scenarios through its Swap tool, in an effort to make projections about the ecological ramifications of landscape change. To use the Swap tool, the user must first identify a covariate to be changed and construct a GIS layer depicting this change. For example, a forecast about the impact of building a new road through a habitat would rely on the construction of a covariate layer that contains the projected road. The user can then apply the fitted RSPF model to this new layer (instead of the original layer), and view the response surface under the changed landscape. We emphasize that such projections are not absolute, they are simply an application of current responses to alternative scenarios, and do not account for potential unobserved threshold values. Furthermore, projections may be faulty if they are made for covariate combinations that never occur in the observed dataset.

The Swap tool resides within the RSPF functionality, and can easily be applied to an RSPF model and surface once projected covariate layers are built. Additional types of alternate landscape conditions include products such as expected forest density after thinning, forage production after burning, or Net Primary Productivity (NPP) under a future climate scenario. An example of the Swap tool is shown in section 9.9

9.0 RSPF Example 1: Pronghorn

9.1 Overview and Narrative Model

Yellowstone National Park pronghorn (*Antilocapra americana)* face a risk of extirpation due to geographic/demographic isolation, low abundance, and low recruitment. Decision makers need a management plan based on demographic monitoring of abundance, especially vital rates and recruitment. This study, led by PJ White, YNP, focused on

1. Demographic monitoring esp. recruitment and survival
2. Ecological interactions esp. predation rates and recruitment

Staging areas, migratory corridors, and summer/winter use area were also of interest here (see Figure 9.1).

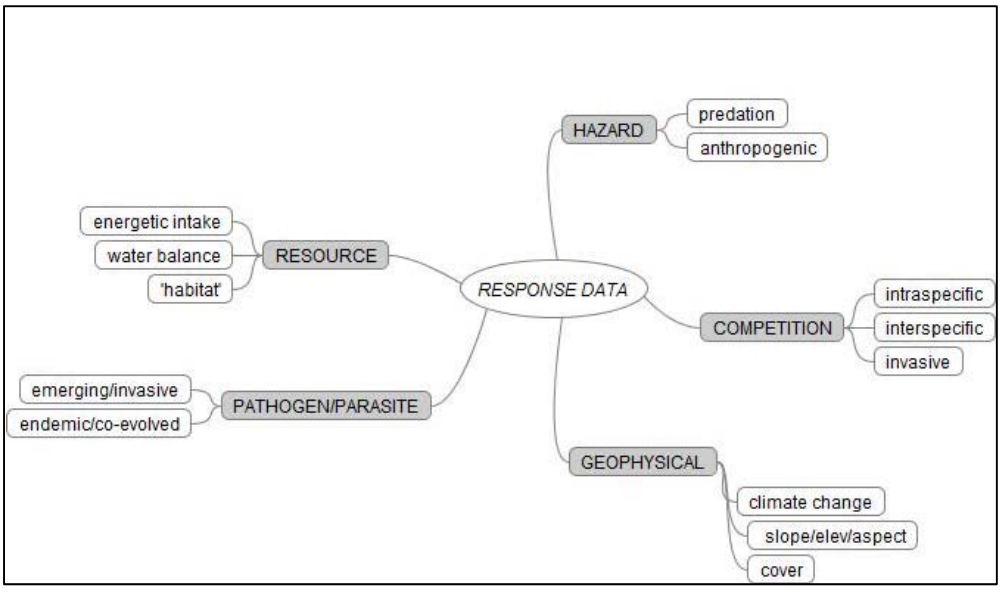


Figure 9.1: Narrative model framework for Pronghorn analysis.

In order to get at a more all-encompassing assessment of vital rates (esp. recruitment), we fit two RSPFs for two responses, one representing selection of birthing arenas (for recruitment-specific analysis) and one representing resource selection in general. Here, we include only the results from the general RSPF analysis.

9.2 Data Inputs

Ideally, we are interested in addressing questions of road impacts, predator impacts, and range condition impacts on pronghorn use and recruitment.

9.2a Covariates

We translated these ecological interested into the following set of covariate layers (see Figure 9.2) to use for model building.

- Abiotic
  - Elevation
  - Slope
  - Topographic complexity
- Biotic: Productivity
  - Forage
  - Net Primary Productivity (NPP)
- Biotic: Landcover
  - Percent forest cover
  - Percent sagebrush cover
  - Percent herbaceous cover
  - Percent soil cover
- Biotic: Predation
  - Coyote intensity of use
  - Wolf intensity of use
  - Small mammal (prey) prevalence
- Human Influenced
  - Distance to roads

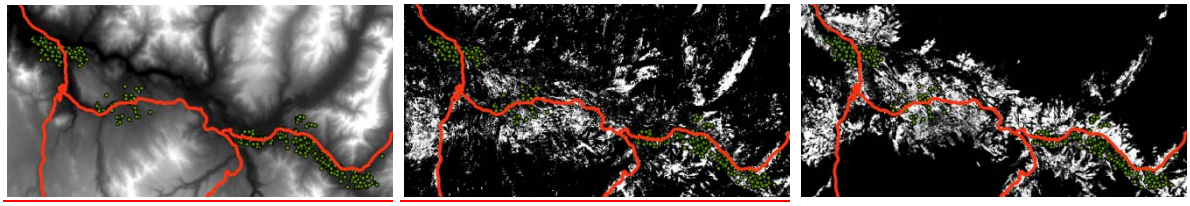


Figure 9.2: Covariate maps for a) elevation, b) forage, c) percent sage

9.2b Model Suite

In order to assess the impact of distance to road on our model, we fit two multi-covariate models, one that included distance-to-road and one that excluded it. To examine predator impacts, we fit a model that excluded coyote and wolf use as predictors, and compared this model to a saturated model, where both coyote and wolf were included. These specific questions led to the following model suite, which were fit and compared using AIC:

Model 1: Saturated model with all covariates fit at an appropriate order
Model 2: Saturated model omitting distance to road
Model 3: Saturated model omitting predators

9.3 Data Integration

Data integration occurred in the ArcGIS environment prior to running the RSPF tool.

9.3a Sampling

Locational data were derived from marked Yellowstone Pronghorn. 762 fixes were made on 26 collared animals from May to July of 2005 of a 1500 km^2 study area (PJ White, Yellowstone National Park ungulate biologist). Figure 9.3a shows a map of the study domain and used locations.

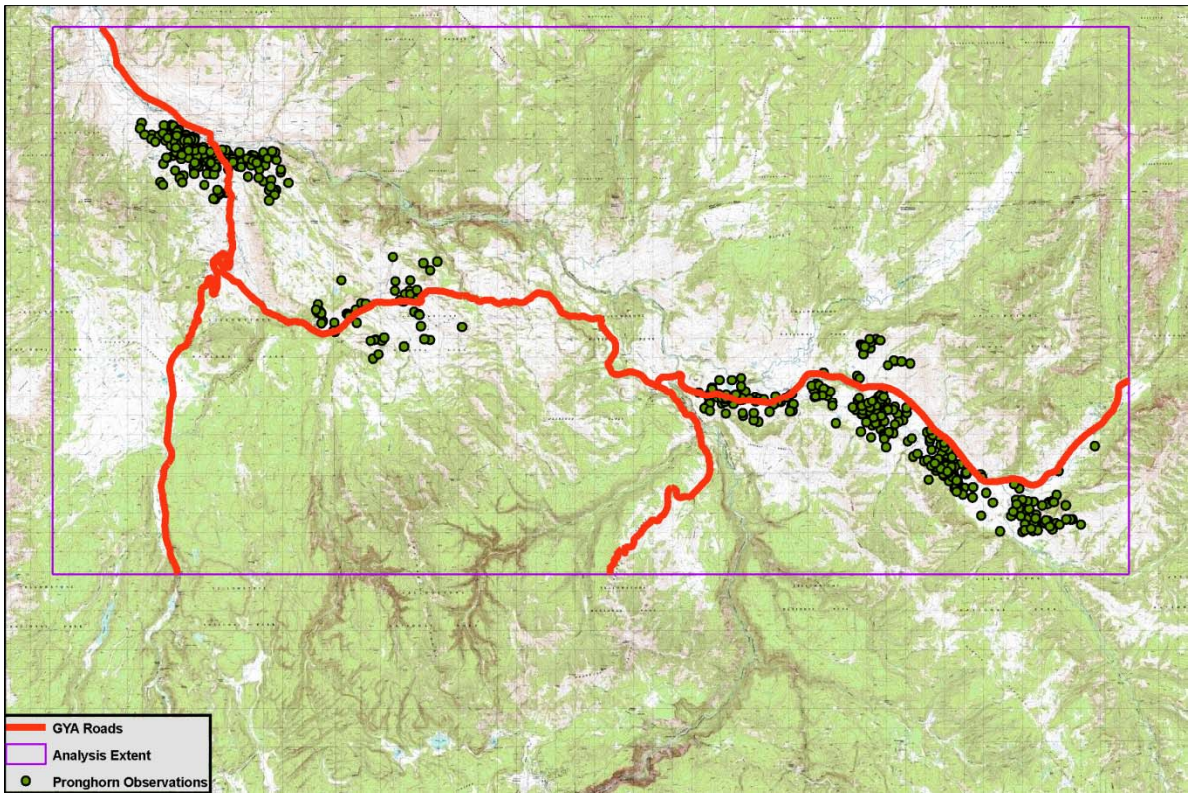


Figure 9.3a: Spatial domain and observed use locations for YNP pronghorn, May-July 2005.

9.3b Full Spatial Domain

Data on pronghorn use were collected on 26 collared individuals from May to July of 2005. After compilation of the pronghorn use data, a full spatial domain encompassing all the use points, as well as some surrounding edge area (to be used for selecting potential available points) was designated. This region was selected arbitrarily by the research team, but was driven in part by the known locations of pronghorn use.

9.3c Modeled Covariates

We used a selection of modeled covariate layers in this analysis. CASA_Forage (YERC) was used to generate the forage layer. Shengli Huang (YERC) generated the herbaceous, sage, and soil layers by modeling AVIRIS satellite imagery and Radar. CASA_Express (YERC) was used for generation of the May and Jun cumulative NPP layers. Small mammal biomass is a modeled layer based on regression of empirically observed biomasses against a habitat map (Alan Swanson, YERC). Coyote and Wolf intensity of use layers were created by accumulating kernel

density surfaces for individual use probabilities to account for pack sizes, and might be considered modeled as well.

9.3d Available Points

Buffers of 1km were generated for all use points to create an "available" space, and available points were randomly and uniformly chosen over that space. Since the spatial scale at which pronghorn select their habitat was unknown, this process was repeated at 3km and 5km, and analyses were conducted at each of these scales for comparative purposes. We arbitrarily selected available points at the 1km scale for this tutorial. Techniques for assessing an optimal scale for availability are in development.

9.3e Spatial Scale of Covariates

All covariates and the response were geo-referenced in the WGS84 UTM zone 12N projected coordinate system. One common pixel size of 100 m grid cells was decided upon, and covariate layers were appropriately up- or down-scaled. Alignment of covariate layers was achieved through resampling. Corners of all grid cells were matched to allow for mapping of the fitted RSPF to the study domain.

9.3f Merged Data Array

A merged data array encompassing used and available points sampled over a common covariate scale was produced in ArcGIS through the RSPF tool, as described in Section 9.4.

9.4 Implementing the RSPF tool

To activate the RSPF tool, the user clicks on the RSPF button adjacent to any open toolboxes in ArcGIS (see Figure 9.4a).

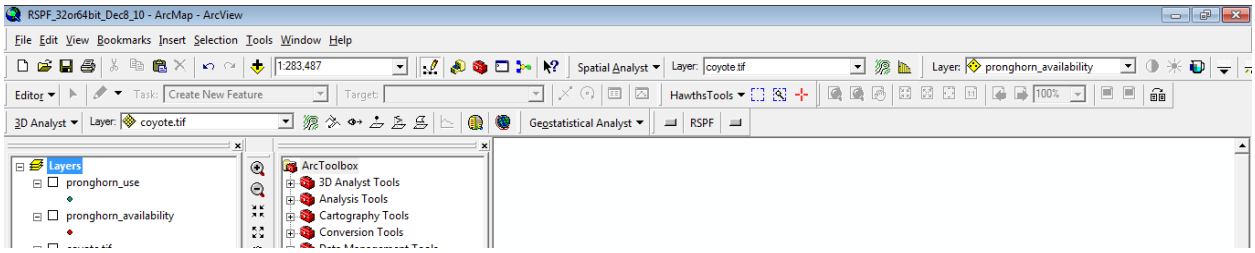


Figure 9.4a: RSPF button displayed in ArcMap.

Upon clicking this button, the screen shown in Fig. 9.4b appears. The user must work through all three tabs prior to submitting their data for analysis. In the first tab, the user must identify the Region of Interest (ROI), which can be any layer that is clipped to the appropriate dimensions. The user must also select a resolution (here, the resolution of the elevation tif, which is 100 m), and identify an output folder where the Run file containing all RSPF output is built.

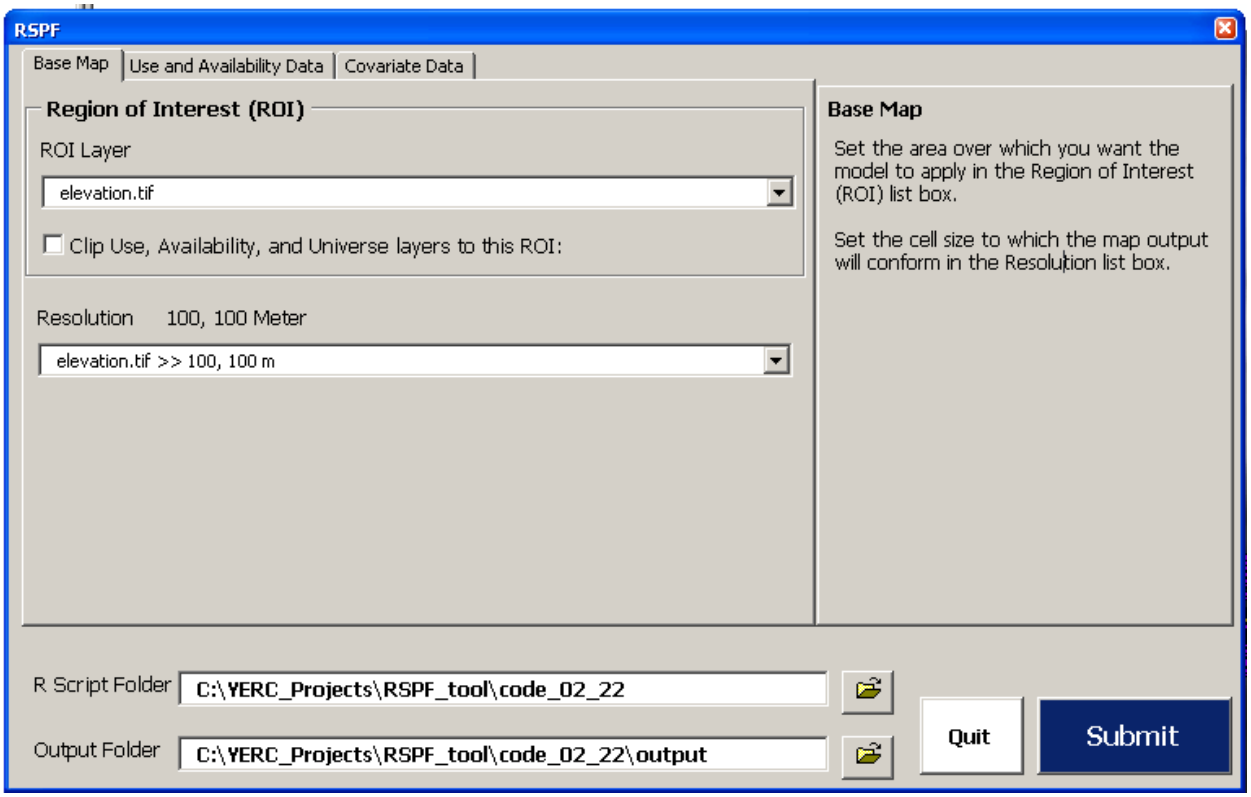


Figure 9.4b: The Base Map tab of the first RSPF user dialog. In this tab, the user enters the ROI layer, sets a file that defines the spatial resolution of analysis, identifies the folder containing the R scripts (i.e., the location of RSPF_script_1.r and RSPF_script_2.r), and the location of the output folder.

In the Response and Availability Files tab (see Figure 9.4b below), the user must identify the layer containing the response measurements (that is, the layer of used points) and select a mechanism for selecting available points (see Section 6.4.2 for descriptions of the mechanisms provided). These mechanisms are represented by the three radio buttons below the Availability File heading. For the pronghorn analysis, we designated a set of points to use for availability,

contained in the Pronghorn Availability shape file, which is selected with the third radio button.

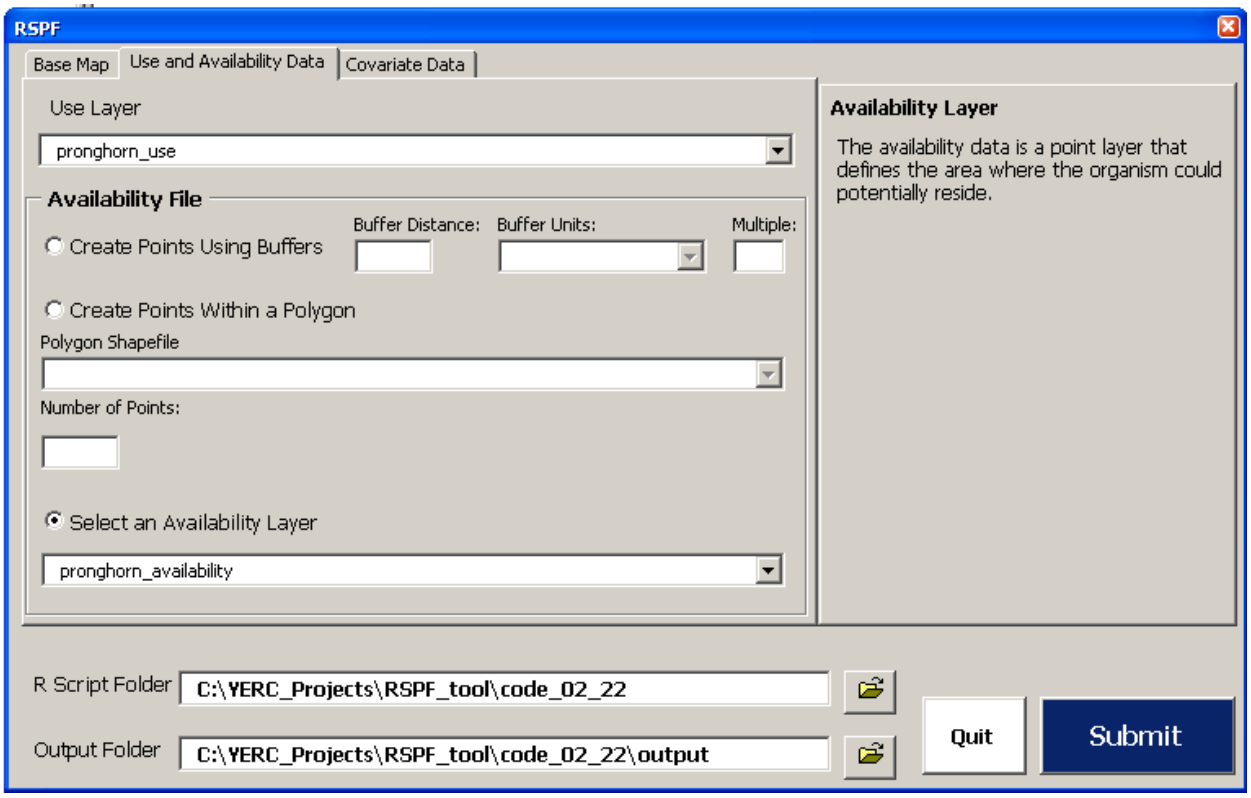


Fig. 9.4c: The Use and Availability Data tab in which the user enters use layer (i.e., the response data) and either makes the availability of specifies a pre-made availability point shapefile.

The third tab of the first user dialogue allows the user to enter all desired covariates for preliminary analysis (see Figure 9.4d). Users can elect to use layers by selecting them from the drop-down menu below Covariate Layers. For the pronghorn analysis, we initially selected all layers for model fitting, as shown below. Selected layers are listed in the large white box below the selection box. At this point, the user can also make several choices about the graphical display of the covariates, by selecting a number of bins and a binning method for the empirical RSPF fit. We selected twelve bins, and bin generation via the quantile method, as we found this generated the most comprehensible picture of the empirical RSPF fit.

Upon completion of all three user dialogue tabs, R is called to fit the univariate RSPF curves by hitting the Submit button in the lower right-hand corner of the user dialogue box.

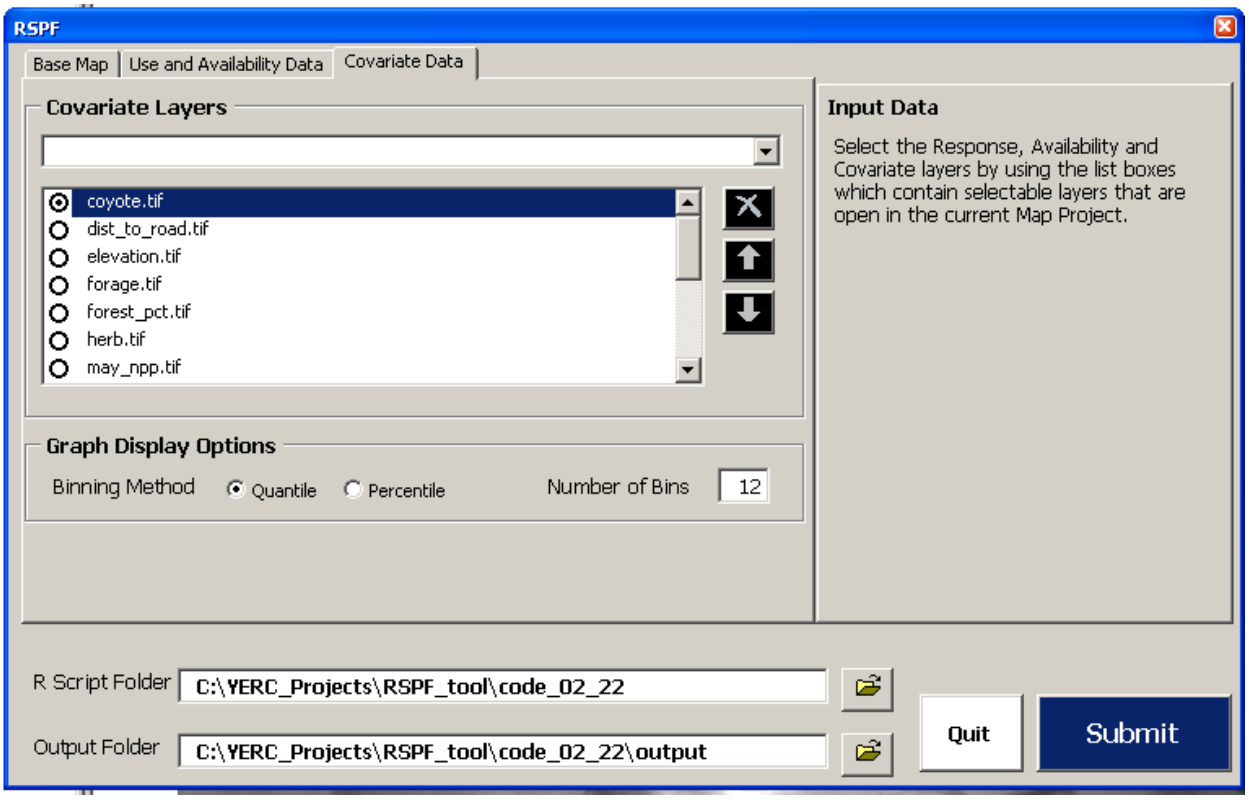


Figure 9.4d: The Covariate Data tab in which the user selects all raster covariates to be included in the analysis. Note that each covariate must be added to the ArcGIS project to be available in this dialog.

A follow-up dialogue box (Figure 9.4e) opens so that the user can designate each covariate as categorical or continuous. At this point, ArcGIS generates a set of random universe points, sampled uniformly over the entire study domain. These points display in ArcGIS, and are used to generate the stacked histograms in R (see Section 9.5). Once extraction of the random universe points is complete, the merged data array is constructed and passed to R (see Figure 9.4f). This script may take several minutes to run, depending on the desired number of points and covariate layers.

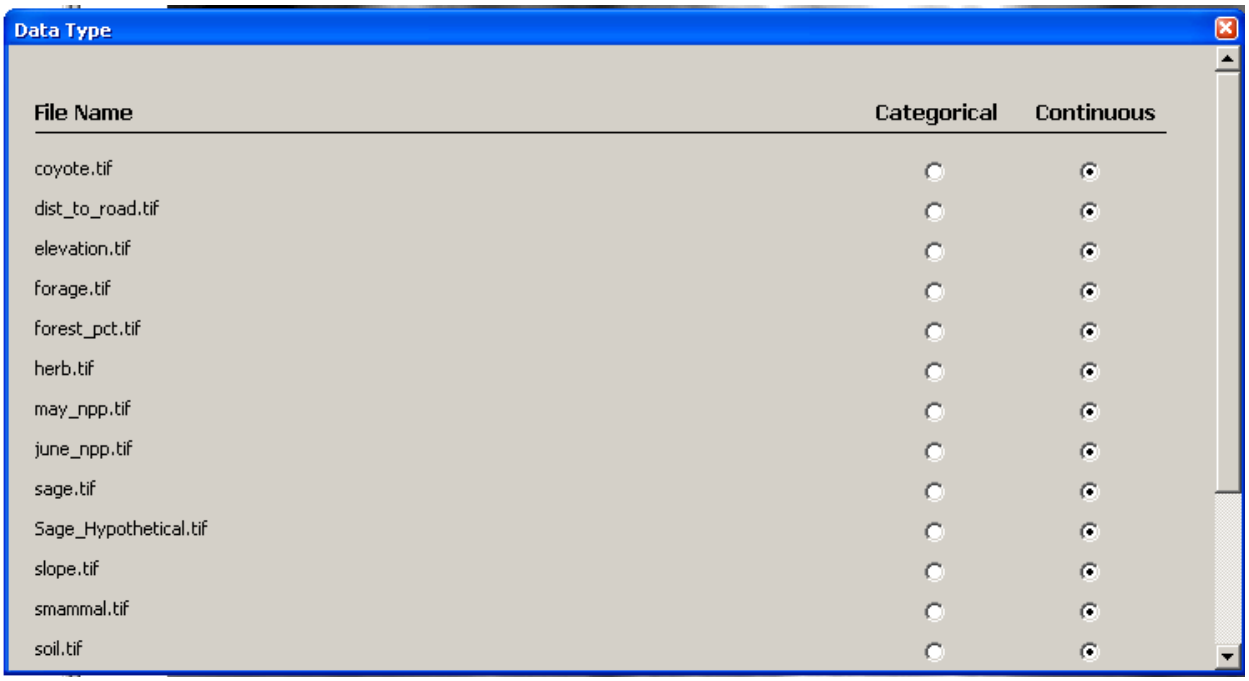


Figure 9.4e: The Data Type selection window in which the users select the appropriate data type for each covariate.

Figure 9.4f: Visualization of merged data array generation through layer stacking.

While the first R script is running, a box will appear on top of the ArcGIS environment, displaying the R output (see Figure 9.4g).

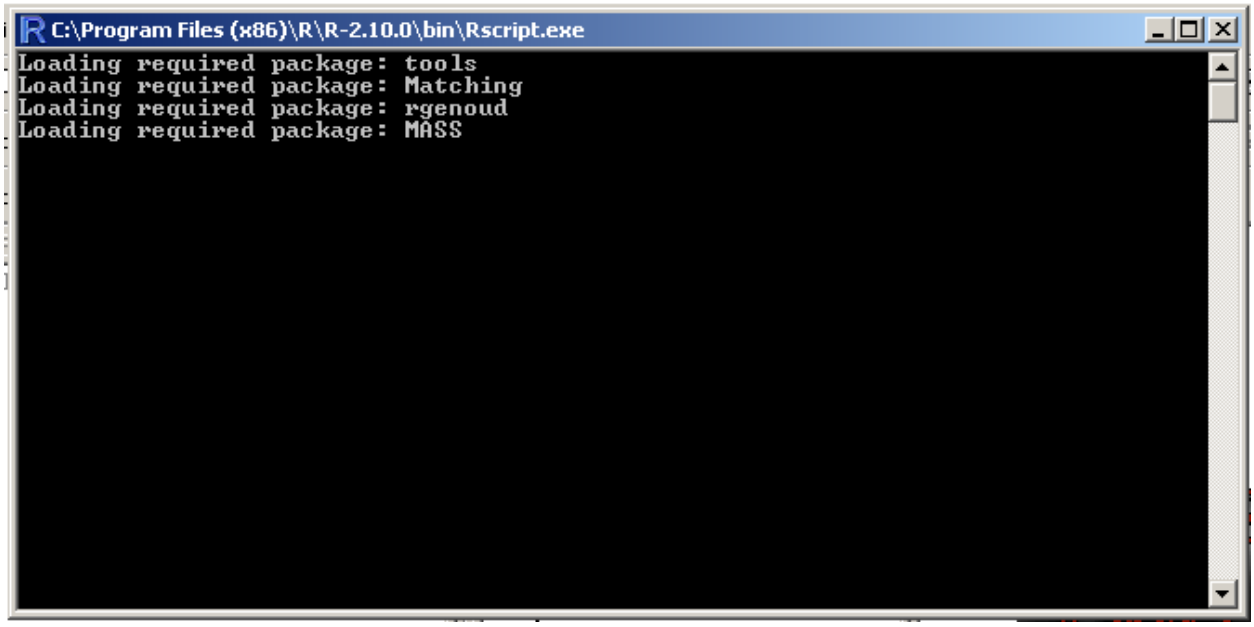


Figure 9.4g: Appearance of ArcGIS when the first R script is running. Note that large amounts of processing are occurring in R while this window displays and a full records of those processes is stored in the file "rspf_log_script1.txt".

9.5 Data Exploration: Boxplots, and Pairsplots for Covariate Distributions

After completion of the first RSPF R script, a display opens in ArcGIS (see Figure 9.5a). This display contains information necessary for data exploration of each covariate, as well as assessment of the link function and covariate term order (first order, quadratic, etc.) for the full RSPF model.

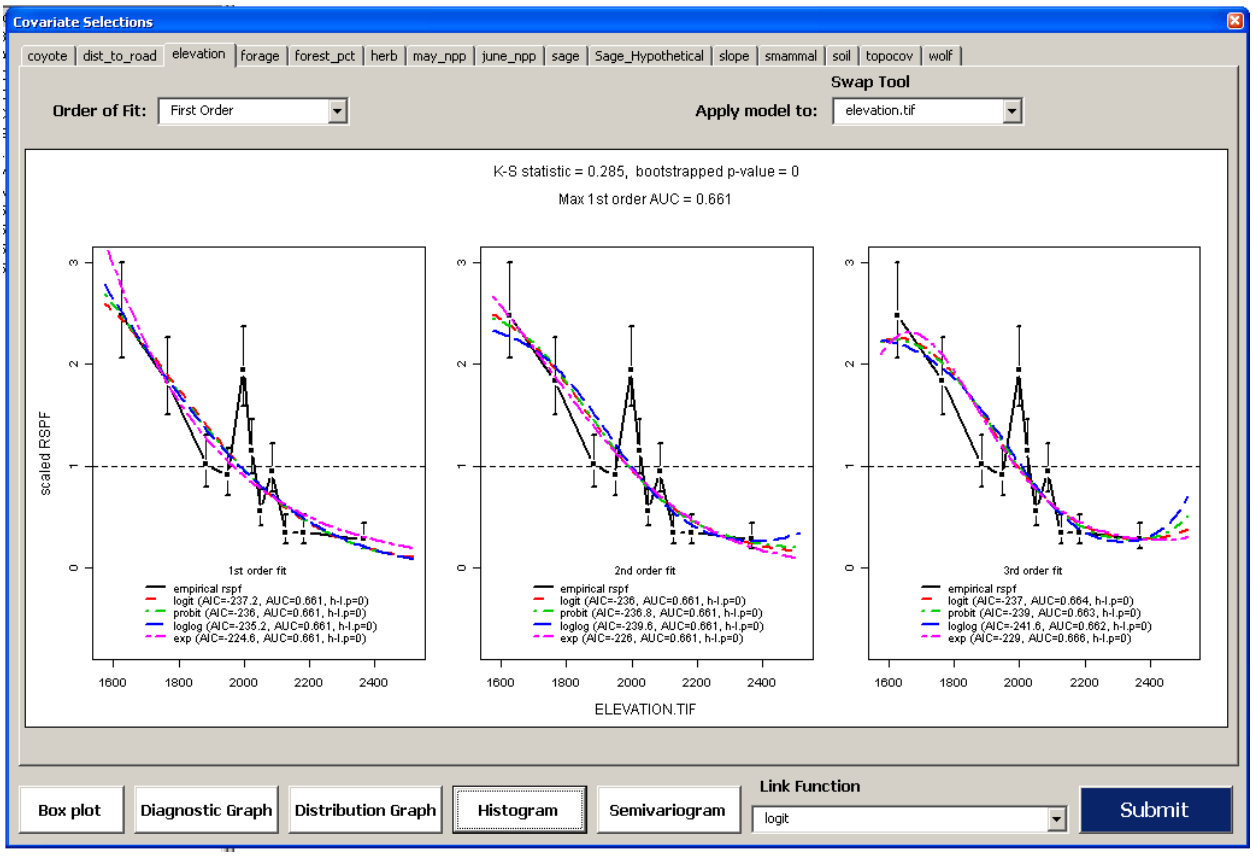


Figure 9.5a: Data displays following the first RSPF R script. The order of fit drop-down menu (upper left) allows the appropriate fit for each variable to be selected. The swap tool drop-down menu allow any variable to be swapped for an alternative dataset, thereby allowing what-if scenarios to be tested (see section 9.7). Clicking the white buttons at the bottom of the window will display the diagnostic graphs for each variable. The link function drop-down menu allows users to select the appropriate link function for their analysis.

Data exploration was conducted in the statistical programming environment R through the ArcGIS shell. A subset of the materials generated in the data exploration is included below.

We examined histograms for each covariate as it occurred in three different cases: cases of Pronghorn Use, cases in the designated Pronghorn Available space, and cases from the entire spatial domain (see Figure 9.5b below). The random universe cases 10000 points distributed uniformly over the entire region of interest. For many covariates, these distributions are similar, but for covariates where the distributions are quite different (for example, for herbaceous cover, shown below), there is some evidence that Pronghorn selection may depend on that covariate.

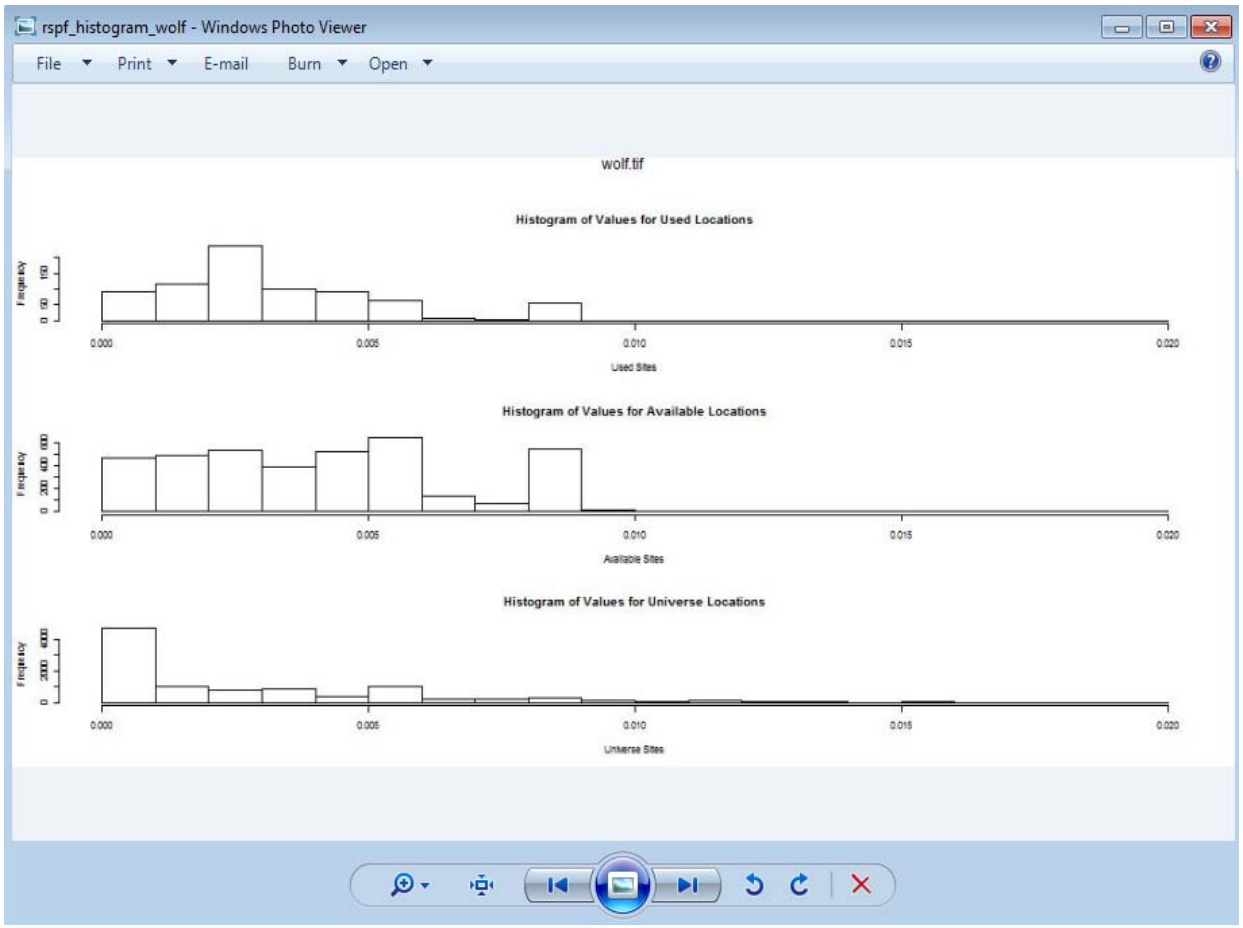


Figure 9.5b: Stacked histograms to compare distributions of universe, used, and available sites.

Boxplots can be used to compare the distributions of different covariates at sites that pronghorn actually used and sites that were deemed available to them. In Figure 9.5c, for wolf intensity of use, we see that the distribution of wolf intensity of use in the dataset is slightly right-skewed (since the boxplot is shifted toward the lower portion of the y-axis), but there do not appear to be substantial outliers in wolf intensity of use.

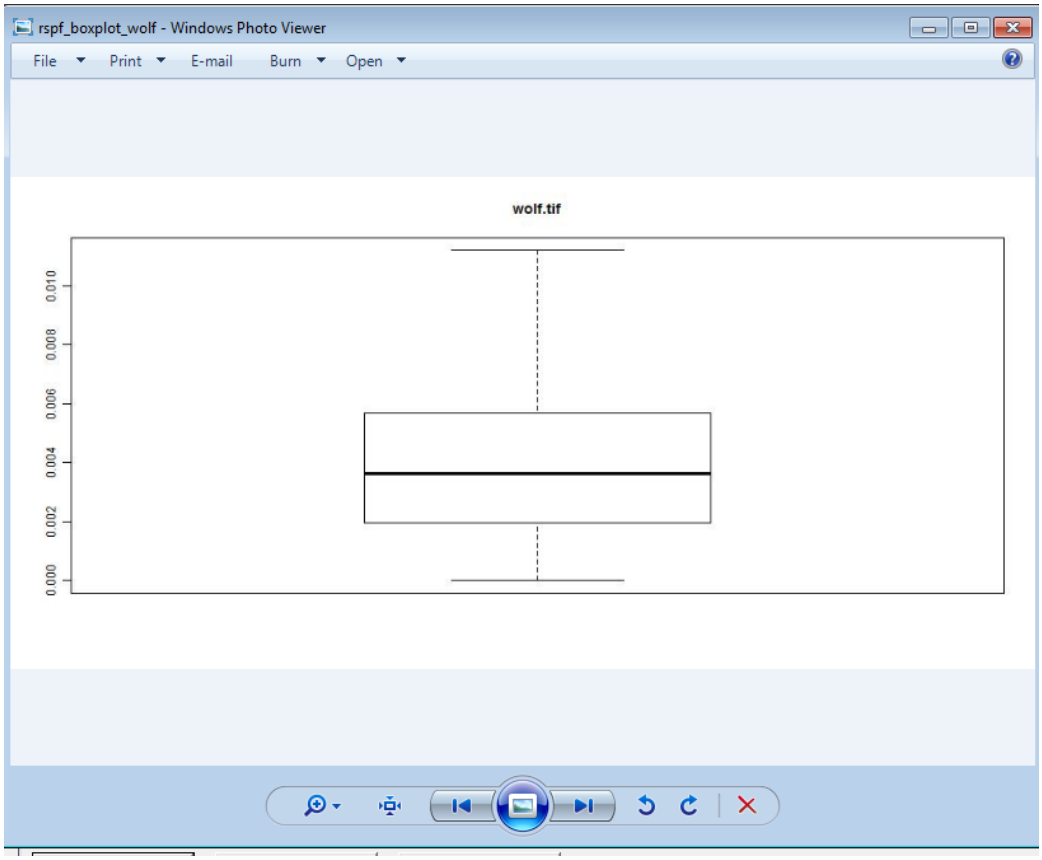


Figure 9.5c: Boxplot of univariate distribution of wolf intensity of use.

A pairs plot (Figure 9.5c) is produced to compare all covariates. This plot matrix is particularly useful in helping researchers identify potentially collinear variables (for example, May and Jun NPP in the pairsplot below). Collinearity is problematic in fitting linear models, thus in general, pairs of collinear variables should not both be included in an analysis.

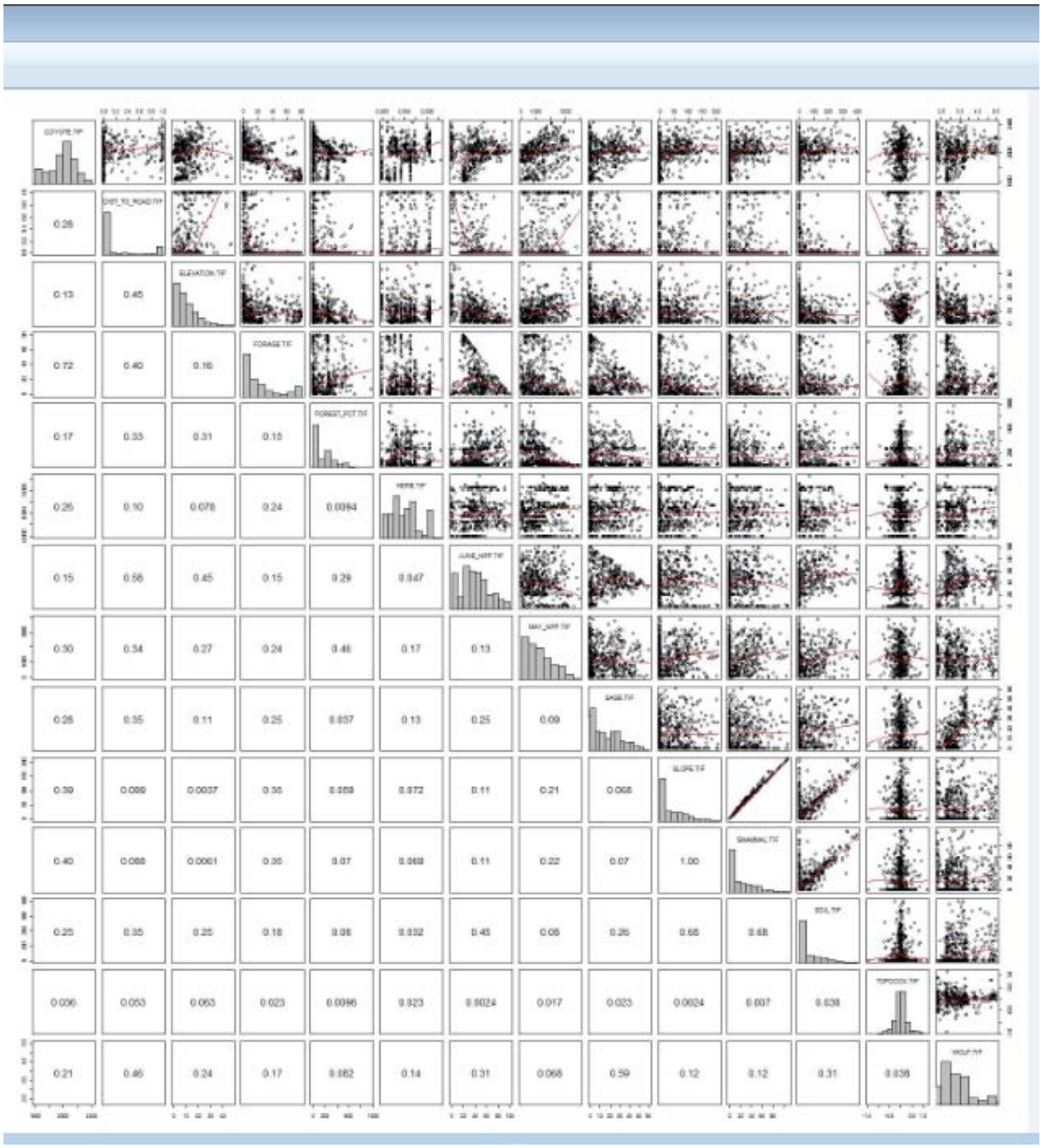


Figure 9.5c: Pairs plot for Pronghorn.

9.6 Assessing Univariate RSPF Curves in ArcGIS

We encourage users to make two passes through the univariate RSPF plots. In the first pass, we recommend focusing on which link function appears to best fit the data. Since only one link function can be chosen for the final model, we are looking for the link function that does the best in general. In the pronghorn example, this appeared to be the logit link. In the second pass

through the plots, we suggest focusing on which order of fit (linear, quadratic, etc.) looked best for each covariate. Here, we focus only on the curves generated by the best link function (in this case, logit). For example, the wolf intensity of use curves shown below (Figure 9.6) are consistent with a linear fit: as wolf intensity of use goes up, pronghorn use declines. Several formal measures of fit are provided for comparison of fits. AIC, our go-to model selection criterion, indicates that the linear fits perform best for wolf intensity of use.

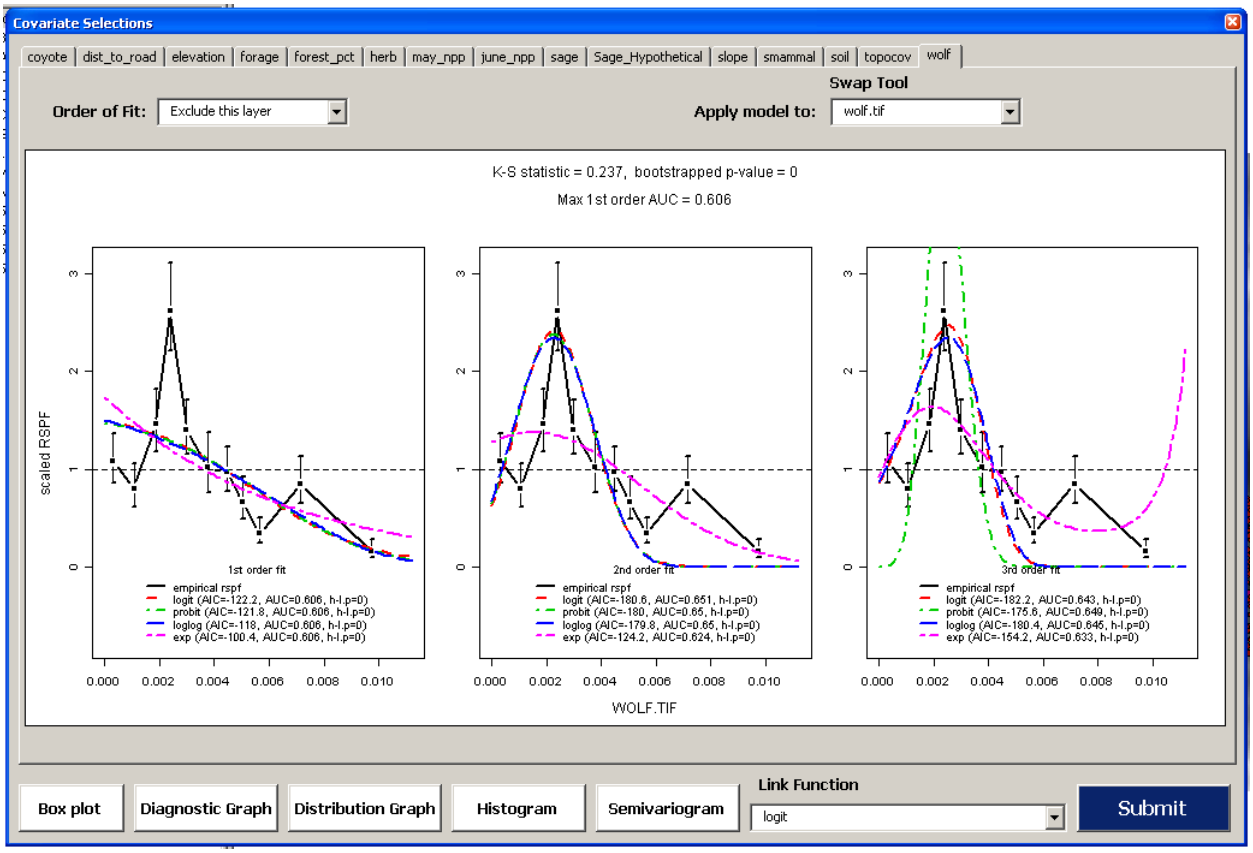


Figure 9.6: Univariate RSPF curve for wolf intensity of use.

9.7 RSPF Example: Second Phase in ArcGIS

The second R script is called after a link function, an order of fit, and an application layer have been selected in each covariate tab of the user dialogue. Send the desired model to R by clicking the Submit button in the lower right-hand corner of the dialogue boxes. The screen will appear to be inactive for several minutes while the second R script runs and the equation is mapped back to the spatial domain. When the fitted RSPF surface appears in the ArcGIS, the second script is complete.

9.8 RSPF Model Selection and Output

Upon completion of the second R script, the RSPF output is stored in the RunX folder located inside the user-designated output directory. The RSPF output is comprised of three parts. First, the ROC curve and semivariogram are created and stored as graphics in the Results subfolder of the RunX file. Second, the RSPF model, as well as AIC and AUC scores, goodness-of-fit tests, and variance inflation factors are stored in the RSPF_summary_file in the Results subfolder of the Run folder. Those results for the saturated Pronghorn model are listed here.

```
n use=762, n avail=3810

Parameter Estimates
                          est     se     t            p  vif
(Intercept)           -12.8088 1.3567 -9.44 4.602682e-20   NA
coyote.tif              0.0821 0.0207  3.96 8.209788e-05  1.4
dist_to_road.tif      -18.8106 2.7598 -6.82 1.876677e-11 72.1
I(dist_to_road.tif^2) -10.5740 1.4697 -7.19 1.571563e-12 71.8
elevation.tif          -0.6405 0.1107 -5.79 1.035356e-08  2.6
forage.tif              0.1094 0.0721  1.52 1.289330e-01  2.9
forest_pct.tif         -0.4398 0.1273 -3.46 5.707549e-04  3.5
herb.tif                0.1928 0.0762  2.53 1.161013e-02  2.5
june_npp.tif           -0.1471 0.1007 -1.46 1.447094e-01  2.4
sage.tif               -0.1954 0.0504 -3.88 1.136533e-04  1.5
slope.tif              -0.4714 0.0783 -6.02 2.730155e-09  1.4
soil.tif               -0.0070 0.0471 -0.15 8.808050e-01  3.0
wolf.tif               -0.5286 0.0642 -8.23 8.288591e-16  1.1

Log-likelihood of GLM estimates:  343.977
Log-likelihood of DC estimates:  NA
Log-likelihood of N-M estimates:  361.8818
AIC of N-M estimates:  -697.7635
AUC for N-M:  0.7666794
mean rspf value for N-M:  0.03986547
Hosmer-Lemeshow goodness of fit results:
chi = 46.7
p = 1.73820593696306e-07
```

A quick assessment of this table illustrates one major problem with this model: the variance inflation factors (VIF) for distance to road and distance to road squared are both quite high, indicating collinearity between those two covariates. A better model would include only a first order distance to road term. Additionally, after careful consideration of the biological ramifications of all covariates considered in the model, the user team determined that June NPP, sage, the two predator covariates (wolf and coyote intensity of use) and forage were unlikely to

be particularly important. Coefficient estimates for a reduced model that is much more interpretable are tabled below.

```
Parameter Estimates
                       est     se      t            p vif
(Intercept)      -11.2101 0.9122 -12.29 8.915039e-32  NA
dist_to_road.tif   0.1842 0.2630   0.70 4.841429e-01 1.2
elevation.tif     -1.3175 0.1362  -9.67 6.228622e-21 1.6
forage.tif         0.0211 0.0422   0.50 6.172208e-01 1.5
forest_pct.tif    -0.8502 0.3207  -2.65 8.217647e-03 2.1
herb.tif           0.6099 0.0945   6.46 1.875431e-10 2.0
slope.tif         -0.4672 0.0843  -5.54 4.179481e-08 1.3
```

Initially, we note that the variance inflation problems present in first model are no longer a problem (variance inflation factors should generally be less than 10, as is true for all covariates in the reduced model).

When interpreting the model coefficients, we remind the user of several important points:

1) All coefficient estimates and covariate significances are based on *all other covariates being in the model*. Thus while in this case we note that the highest individual coefficient significance is attributed to the elevation covariate in the reduced model, that covariate's presence may not actually contribute greatly to the model at large. In order to determine whether a covariate contributes substantially to a model's fit, we recommend fitting models with and without the covariate of interest, and comparing those models' AIC scores, as outlined below.
2) We recommend that the user examine each coefficient's sign and determine whether the sign of the coefficient makes sense (for example, here we see a negative sign on the coefficient for elevation, and it makes sense that as elevation increases, use by pronghorn should probably decrease, so we are satisfied with that value). *If coefficients' signs are not what is expected, consider fitting a model without that covariate, and comparing model performance (via AIC) to see if inclusion of the covariate is appropriate.*
3) We remind the user that all first-order quantitative covariates were standardized prior to fitting, thus coefficient magnitudes are in terms of standard deviations above or below that covariate's mean value.

4) We suggest considering models that exclude insignificant predictors (e.g.,distance to road in our reduced model). However, we recommend that the user consult Hosmer-Lemeshow p-values or AIC values for models with and without the covariate in order to help decide whether covariate conclusion is appropriate. It is advisable to keep insignificant covariates *if* the sign associated with them makes good biological sense. In general, proximity to roads seems to facilitate animal use in the Lamar Valley (based on RSPFs for several other species), so the positive sign on the coefficient here is unexpected, and removing distance to road from the model might be a prudent choice.

To interpret the elevation coefficient from the reduced model above, one could say that for each standard deviation of increase in elevation, the probability of use by pronghorn decreases by exp(-1.3175) = .268, or 26.8%, at the mean level of all the other covariates included in the model. Similarly, to interpret the model coefficient for herbaceous cover (herb.tif), one could say that for each additional standard deviation increase in herbaceous cover, the probability of use by pronghorn increases by exp(.6099) = 1.84 or 184%, at the mean level of all other modeled covariates.

A ROC plot is located in the Run folder, in the RSPF_ROC_Semivariogram file. The ROC plot for the saturated pronghorn model is shown below in Figure 9.8. The ROC plot here suggests that the model is doing a fairly good job of classifying points as Used or Available.

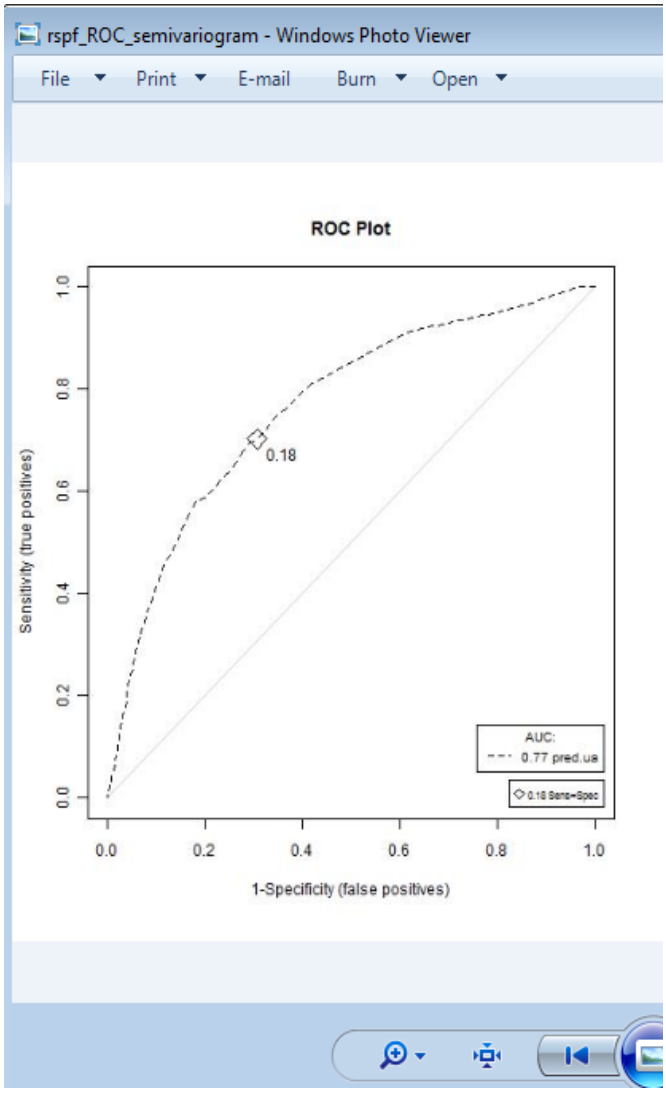


Figure 9.8: ROC plot for the saturated pronghorn model.

This model's AUC is fairly high (AUC = .77), suggesting that the model does a pretty good job of correctly classifying used and available points. The Hosmer-Lemeshow goodness-of-fit test indicates a significant lack of fit in this model, suggesting potential omission of important covariates. However, we are not particularly concerned with the lack of fit, since our objective is to predict with this model, and its AUC is high.

To examine the ecological impacts of distance to road and predation on pronghorn habitat use, we compared AIC scores from our (original) saturated model and two reduced models (one excluding distance to road and one excluding predators, see Section 9.2b). The reduced models

were both fit in their own runs of the RSPF R scripts. AIC values for each of the models is reported here, along with the number of parameters in the model (k), the difference in AIC scores between the best model and this particular model (Δ AIC), and the AIC weight (ω) attributed to that model.

Model	AIC Score	k	Δ AIC	ω
Saturated	-697.7635	13	0	1
No Road	-613.8171	12	-83.95	5.91e-19
No Predators	-605.9269	11	-91.84	1.14e-20

Based on these AIC values tabled above, we conclude that the saturated model performs best (with virtually no weight placed on the other two models in the suite), thus there is a strong indication that pronghorn are responding to both road and predators, when all other covariates are included in the model. To address the road impact question, we fit a model without roads, and compared it to a model that included roads. The road model was superior based on AIC (-614 for the no-roads model, as compared to -698 for the saturated model). To address the predation question, we compared models with and without wolf and coyote. In this case, the saturated model out-performed the model without predators (AIC of the saturated model was -698; for the model without predators it was -606), which suggests that wolf and coyote intensity of use do drive pronghorn resource selection.

The final component of the RSPF output is the predicted RSPF surface for the best model, which is fitted and displayed in ArcGIS (see Figure 9.8). This prediction looks reasonable based on biological knowledge of this system: The large swatch of good habitat that is apparently not used in the upper left-hand corner of the surface is a private in-holding

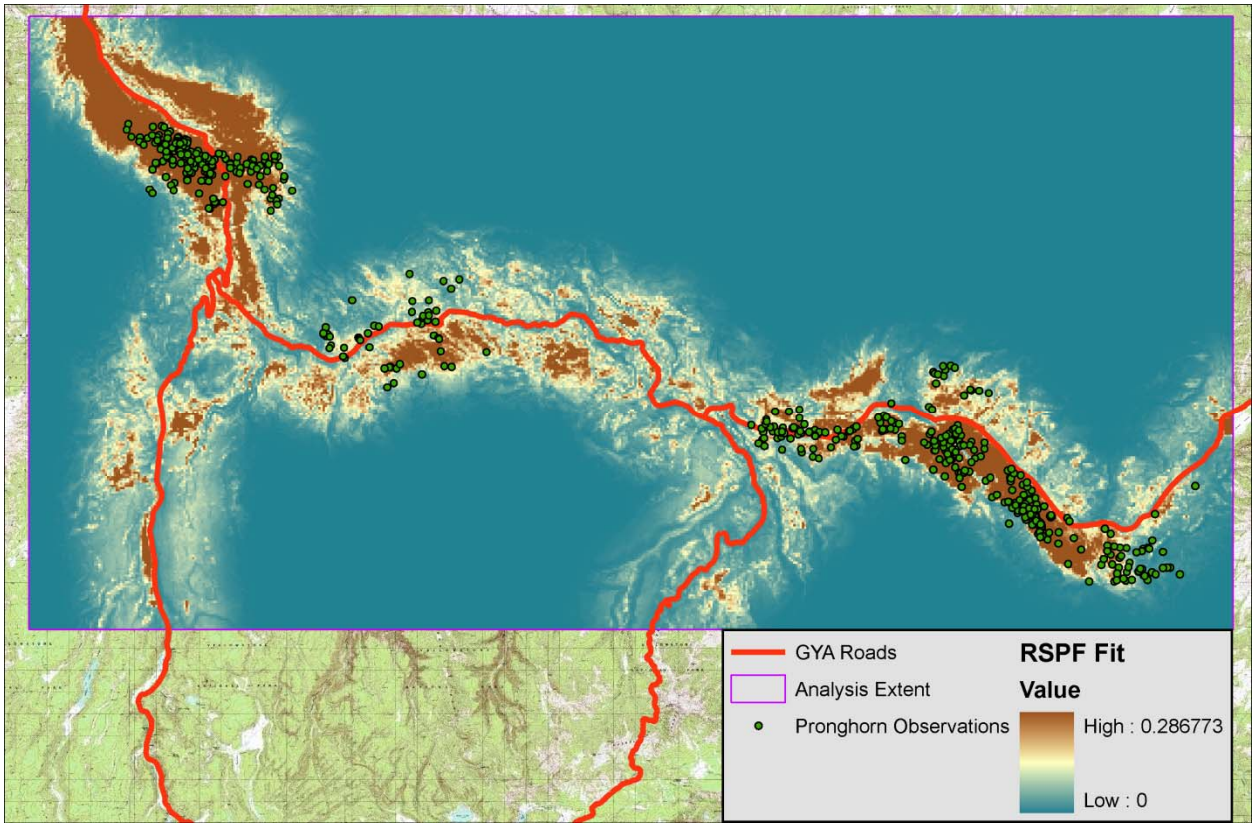


Figure 9.8: RSPF surface as fitted by the final model.

9.9 Scenario Testing

We used two hypothetical scenarios to demonstrate how the EAGLE tool might be used to assess the potential impact of landscape change on pronghorn distribution. These hypothetical "what-if" scenarios were tested using the Swap tool whereby a variable used to define the model (e.g., sage and/or distance_to_road) is replaced with a hypothetical variable when the model is applied (e.g., sage_hypothetical and/or distance_to_road_hypothetical) (Figure 9.9a). RSPF fit results from a model with and without swapping the variable distance to road are shown in Figure 9.9b). These types of What-if-Scenario (WIS) will provide practitioners with important decision support to guide site-level action plans, restoration efforts, and understand the environmental impacts from climate disruptions, invasive species, changing land-use, and disturbance regimes.

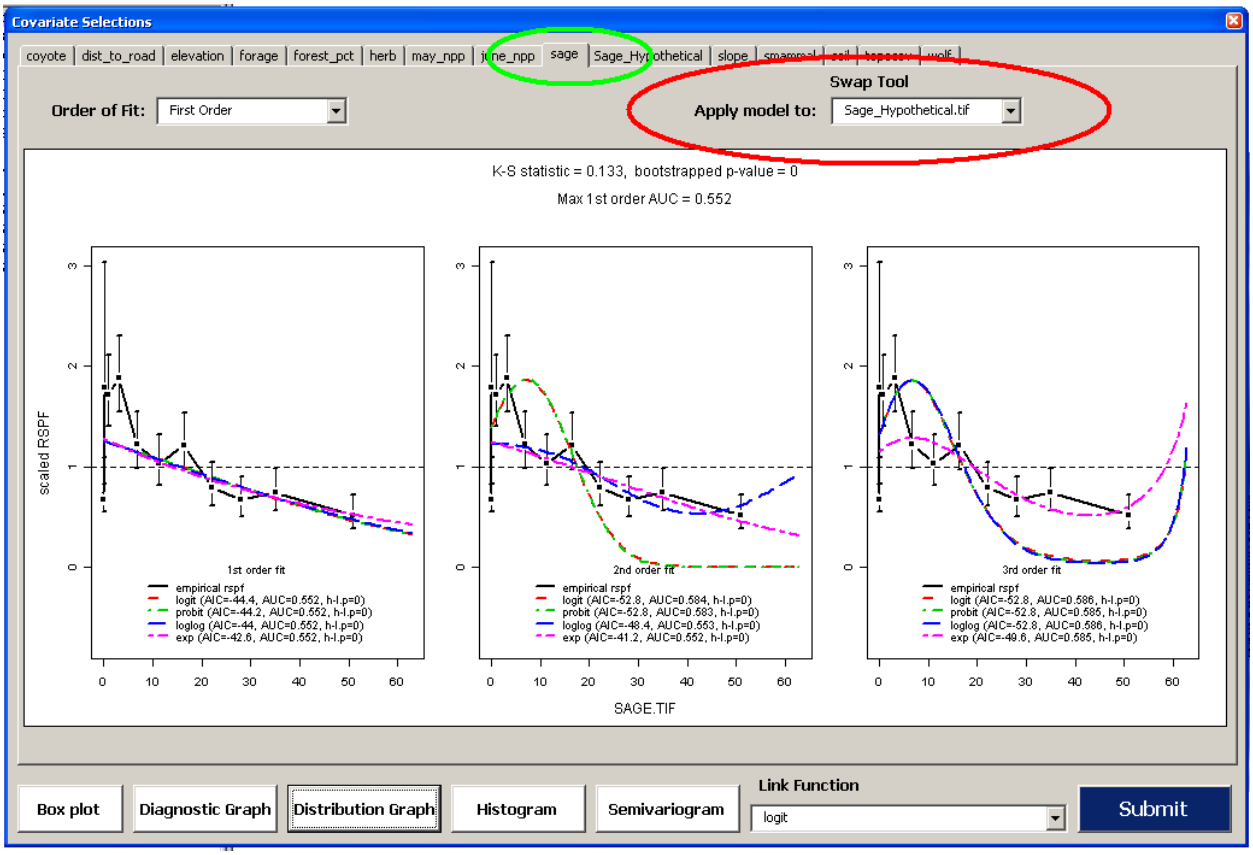


Figure 9.9a: The Swap tool applied to the variable percent sage cover (see green oval). The RSPF tool defaults to applying the RSPF model to the same variable upon which it was built, but the Swap tool allows users to direct ArcGIS to apply the model to an alternative version of that variable (see red oval).

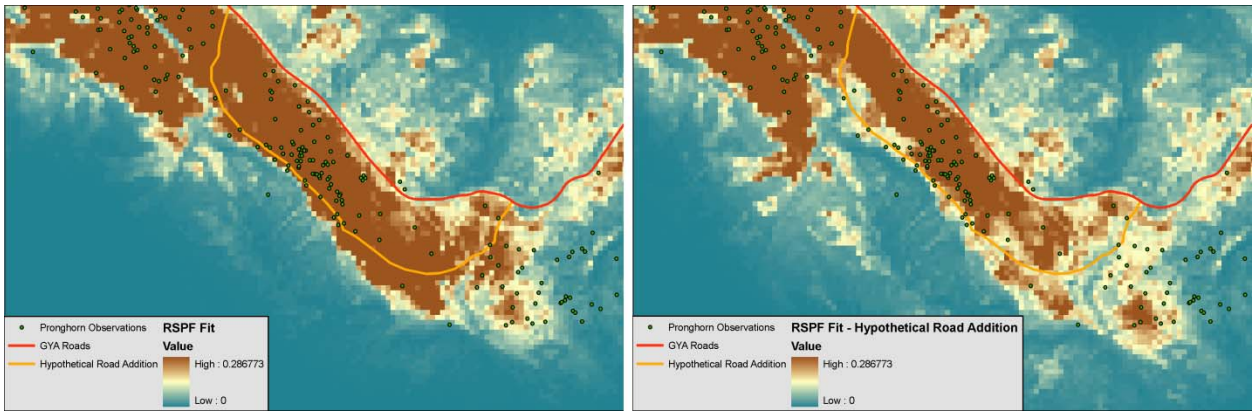


Figure 9.9b: A portion of the original RSPF model output indicating the resource selection function for pronghorn in Yellowstone National Park (left). The Swap was used to apply the RSPF model to an alternative distance to road layer created using a hypothetical road addition (shown in orange). The new prognostic RSPF model output for pronghorn (right) indicates that pronghorn are excluded from portions of their original selected habitats.

10.0 Acknowledgements, Literature, and Programs Cited, Further Readings, and Citation Information

10.1 Acknowledgements

Yellowstone Ecological Research Center thanks the following scientists for their generous contributions to the EAGLES System. The team responsible for creating EAGLES includes:

Robert Crabtree and Jennifer Sheldon (YERC) – Project design
Subhash Lele (University of Alberta) – Statistical consultation
Alan Swanson (YERC, University of Montana) – Project design and R programming (Phase 1)
John Shupe (NASA-Ames) – ArcGIS interface programming (VB) (Phase 1)
Brandt Winkelman (YERC) – Project testing (Phase 1)
Daniel Weiss (YERC) - Project design, management, and testing (Phase 2); COASTER
Gordon Reese (Colorado State University) - ArcGIS interface programming (VB) (Phase 2)
Kezia Manlove (YERC) - R programming (Phase 2)
Aaron Doern (YERC) – GeoSpatial Data Wiki project design and implementation
Katie Gibson – COASTER web programming

Additionally, we greatly appreciate the support and ongoing feedback provided by our early adoption group, especially Greg Watson, Phillip Martin, Jennifer Jenkins, Sean Finn, Sharon Baruch-Mordo, Scott Bergen, Matt Holloran, James Broska, Pat Heglund, and Kurt Johnson. Thanks also to the attendees of the May, 2010 RRSC workshop, including Mark Bertram, Donna Brewer, Stephen DeStefano, Rex Johnson, James Forester, Jonna Katajisto, Jonah Keim, Paul Moorcroft, Doug Ouren, Lori Pruitt, Tara Wertz, Ken Wilson, and Joe Witt, for their feedback and interest. This project was supported by funding from NASA Ecological Forecasting – RRSC - NASA Grant no. NNX08AO58G

10.2 Literature Cited

Beel, J., B. Gipp, and C. Müller. (2009). "'SciPlore MindMapping' – A Tool for Creating Mind Maps Combined with PDF and Reference Management." D-Lib Magazine, 15(11).

Braunisch, V. and R. Suchant. (2010). "Predicting species distributions based on incomplete survey data: the trade-off between precision and scale." Ecography (33): 826-840.

Burnham, K. and D. Anderson. (2002). Model selection and multi-model inference: a practical information-theoretic approach. Second Edition. Springer-Verlag. New York.

Elith, J., C.H. Graham et al. (2006). "Novel methods improve prediction of species' distributions from occurrence data." Ecography 29: 129-151.

Forester, J., H. I., and P. Rathouz. (2009). **"**Accounting for animal movement in estimation of resource selection functions: sampling and data analysis". *Ecology* Vol. 90(12), pp. 3554-3565.

Friedman, J.H. (1991). "Multivariate adaptive regression splines". The Annals of Statistics Vol. 19 (1): 1-141.

Lele, S. and J. Keim. (2006). "Weighted distributions and estimation of resource selection probability functions". *Ecology* Vol. 87(12), pp. 3021-3028.

Guisan, A. and N.E. Zimmermann. (2000). "Predictive habitat distribution models in ecology." Ecological Modeling 135: 147-186.

Lele, S. (2009). "A new method for estimation of resource selection probability function". *The Journal of Wildlife Management* Vol. 73(1), pp. 122-127.

Manly, B., L. McDonald, D. Thomas, T. McDonald, and W. Erickson. (2002). Resource selection by animals. Second Edition. Kluwer Academic Publishers, Dordrecht, Netherlands.

Nelder, J. and R. Mead. (1965). "A Simplex Method for Function Minimization". *Computer Journal* Vol. **7,** pp. 308-313.

Phillips, S.J., R.P. Anderson and R.E. Schapire. (2006). "Maximum entropy modeling of species geographic distributions." Ecological Modeling 190: 231-259.

Sciplore. http://www.sciplore.org/software/sciplore_mindmapping/

Smithson, M. and J. Verkuilen. (2006). "A Better Lemon-Squeezer? Maximum-likelihood regression with beta-distributed dependent variables." Psychological Methods Vol. 11(1): 54-71.

Zuur, A,F., E.N. Ieno and G. Smith. (2007). Analyzing Ecological Data. Springer Science, New York.

Zuur, A.F., E.N. Ieno and C.S. Elphick. (2010). "A protocol for data exploration to avoid common statistical problems." Methods in Ecology and Evolution (1): 3-14.

10.3 R Packages Used and Citations

These are the packages required by the RSPF tool that are NOT included in the standard installation of R. Note that all of these packages must be added to the R library folder as specified in section 2.1b. All necessary files are provided in the .zip file called RSPF_R_libraries.zip.

geoR
HH
leaps
Matching
MCPAN
multcomp
mvtnorm
PresenceAbsence
rgenoud
sp

Freeman, Elizabeth (2007). PresenceAbsence: An R Package for Presence-Absence Model Evaluation. USDA Forest Service, Rocky Mountain Research Station, 507 25th street, Ogden, UT, USA.

Heiberger, Richard M. (2009). HH: Statistical Analysis and Data Display: Heiberger and Holland. R package version 2.1-29.

Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2), http://cran.r-project.org/doc/Rnews/.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ribeiro, Paulo J. & Peter J. Diggle (2001). geoR: a package for geostatistical analysis R-NEWS, 1(2):15-18. June, 2001.

Schaarschmidt, Frank, Daniel Gerhard and Martin Sill (2008). MCPAN: Multiple comparisons

using normal approximation. R package version 1.1-7.

10.4 Further Readings

Miller, J., J. Franklin and R. Aspinall, (2007). "Incorporating spatial dependence in predictive vegetation models". Ecological Modeling 202: 225-242.

Legendre, P. and L. Legendre, (1998). Numerical Ecology. Elsevier, Amsterdam.

10.5 Citation Information

Manlove, K.R., Weiss, D.J., and Sheldon, J.W. (2011). EAGLE User Manual. Yellowstone Ecological Research Center. Bozeman, MT.

Appendix 1: List of Covariate Layers Commonly Used by YERC

Description	Dataset Name	Resolution	Frequency	Period
Maximum Temperature	TOPS - climate	1km	Daily	from 1950
Minimum Temperature	TOPS - climate	1km	Daily	from 1950
Precipitation	TOPS - climate	1km	Daily	from 1950
Short Wave Solar Radiation	TOPS - climate	1km	Daily	from 1950
Vapor Pressure Deficit	TOPS - climate	1km	Daily	from 1950
Dew Point Temperature	TOPS - climate	1km	Daily	from 1950
GPP/NPP, Gross and Net Productivity	TOPS - modeled	1km	Daily	from 2000
Snow	TOPS - modeled	1km	Daily	from 2000
Evapotranspiration	TOPS - modeled	1km	Daily	from 2000
Outflow	TOPS - modeled	1km	Daily	from 2000
Soil Moisture	TOPS - modeled	1km	Daily	from 2000
Phenology	TOPS - modeled	1km	Daily	from 2000
Vegetation Stress	TOPS - modeled	1km	Daily	from 2000
Snow Cover	MODIS	500m	8 day	from 2000
Land Surface Temperature & Emissivity	MODIS	1km	8 day	from 2000
Vegetation Indices (EVI, NDVI)	MODIS	1km	16 day	from 2000
LAI and FPAR	MODIS	1km	8 day	from 2000
GPP	MODIS	1km	8 day	from 2000
Land Cover Type	MODIS	1km	Yearly	from 2000
Thermal Anomalies and Fire	MODIS	1km	8 day	from 2000
Albedo	MODIS	1km	8 day	from 2000

GPP Gross and Net Primary Productivity	MODIS	5km	daily	from 2000
FPAR Fraction of Photosynthetically Active Radiation	AVHRR	8km	8 day	from 1982
Net Primary Productivity (NPP)	CASA_Express	2m to 250m	Daily	from 1980
Potential Evapotranspiration (PET)	CASA_Express	2m to 250m	Daily	from 1980
Soil Moisture—3 layers to root depth	CASA_Express	2m to 250m	Daily	from 1980
Snow Water Equivalent (SWE)	CASA_Express	2m to 250m	Daily	from 1980
Herbaceous (Foliar) Biomass Production -- upland	CASA_Forage	2m to 250m	Daily	from 1980
Herbaceous (Foliar) Biomass Production -- wetland	CASA_Forage	2m to 250m	Daily	from 1980
Woody Shrub Biomass Production -- sagebrush	CASA_Forage	2m to 250m	Daily	from 1980
Woody Shrub Biomass Production -- wetland	CASA_Forage	2m to 250m	Daily	from 1980
Stream and River Discharge	CASA_Hydra	250m	Monthly	from 2000
Snowmelt Rate	CASA_Hydra	250m	Monthly	from 2000
Water Temperature in Rivers and Lakes	CASA_Hydra	250m	Monthly	from 2000
Dissolved Oxygen	CASA_Hydra	250m	Monthly	from 2000
Growing Season Length (in days)	CASA_Hydra	250m	Annual	from 2000
Drought – User Specified and Probabilistic	CASA_Hydra	250m	Annual	from 2000
Urban Expansion	PSI/NASA/Minnesota	250m	Annual	from 2000
Agriculture Expansion – New Irrigated Cropland	PSI/NASA/Minnesota	250m	Annual	from 2000
Agriculture Expansion – CRP for two years or more	PSI/NASA/Minnesota	250m	Annual	from 2000
Wetland Conversion to cropland	PSI/NASA/Minnesota	250m	Annual	from 2000
Wetland Loss (drained or dried out)	PSI/NASA/Minnesota	250m	Annual	from 2000
Wetland Expansion	PSI/NASA/Minnesota	250m	Annual	from 2000
Fires (non-forest)	PSI/NASA/Minnesota	250m	Annual	from 2000
Fires (forested)	PSI/NASA/Minnesota	250m	Annual	from 2000
Insect Kill (forested)	PSI/NASA/Minnesota	250m	Annual	from 2000
Logging (forested)	PSI/NASA/Minnesota	250m	Annual	from 2000
AM Freeze Thaw	University of Montana	25km	daily	from 1988
PM Freeze Thaw	University of Montana	25km	daily	from 1988
Combined Freeze/Thaw	University of Montana	25km	daily	from 1988
Inverse Transitional Freeze/Thaw	University of Montana	25km	daily	from 1988
Minimum Temperature	PRISM	4 km	Monthly	1895 to 2010

Maximum Temperature	PRISM	4 km	Monthly	1895 to 2010
Average Temperature	PRISM	4 km	Monthly	1895 to 2010
Percent Normal Precipitation	PRISM	4 km	Monthly	1895 to 2010
Percent Surface Water (PSW)	YERC	1km	8 day	from 2000
Percent Soil, Herbaceous, Shrub	YERC	30m	static	static
Forest Biomass	YERC	100m	Annual	from 2005
Riparian (hydrologically influenced soil) vs. Upland	YERC	30m	Annual	from 1980
Grey Attack, Red Attack, Healthy Green (Forest)	YERC	2m	Annual	from 1980
Annual Average Precipitation	BioClim	1 km	Yearly	1980 to 1997
Annual Average Temperature	BioClim	1 km	Yearly	1980 to 1997
Annual Temperature Range	BioClim	1 km	Yearly	1980 to 1997
Average Diurnal Range In Temperature	BioClim	1 km	Monthly	1980 to 1997
Average Temperature of Coldest Quarter	BioClim	1 km	Quarterly	1980 to 1997
Average Temperature of Driest Quarter	BioClim	1 km	Quarterly	1980 to 1997
Average Temperature of Warmest Quarter	BioClim	1 km	Quarterly	1980 to 1997
Average Temperature of Wettest Quarter	BioClim	1 km	Quarterly	1980 to 1997
Mean Diurnal Range/Annual Temperature Range	BioClim	1 km	Yearly	1980 to 1997
Maximum Temperature of Warmest Month	BioClim	1 km	Monthly	1980 to 1997
Minimum Temperature of Coldest Month	BioClim	1 km	Monthly	1980 to 1997
Precipitation of Coldest Quarter	BioClim	1 km	Quarterly	1980 to 1997
Precipitation of Driest Month	BioClim	1 km	Monthly	1980 to 1997
Precipitation of Driest Quarter	BioClim	1 km	Quarterly	1980 to 1997
Precipitation Seasonality	BioClim	1 km	Seasonal	1980 to 1997
Precipitation of Warmest Quarter	BioClim	1 km	Quarterly	1980 to 1997
Precipitation of Wettest Month	BioClim	1 km	Monthly	1980 to 1997
Precipitation of Wettest Quarter	BioClim	1 km	Quarterly	1980 to 1997
Temperature Seasonality	BioClim	1 km	Seasonal	1980 to 1997
National Land Cover Database 1992	NLCD 1992 (landsat)	30m	Single year	1992
National Land Cover Database 2001	NLCD 2001 (landsat)	30m	Single year	2001
U.S. General Soil Map	STATSGO	Polygons	Single year	2006
DEMs (slope, aspect, elevation, etc.)	Digital Elevation Model	30m	static	unknown
DRGs (distance metrics, road density, stream density)	Digital Raster Graphics	n/a	static	unknown

Globally Downscaled Climate Projections	Future Climate Grids	17 km	Past & future	1961 to 1990, 2041 to 2060, 2081 to 2100

Appendix 2: Specific R Functions Used for Each Model

glm (package nlme)

ROC curve (package PresenceAbsence)

Appendix 3: RSPF Flow of Control Overview

The RSPF tool merges the spatial analytical capabilities of ArcGIS with the statistical functionality available in the R software package. The key components of the RSPF tool are an ArcGIS Graphical User Interface (GUI; written in Microsoft Visual Basic 6.5) and two R scripts that are called by ArcGIS and execute in a DOS-shell. The general flow of control within the RSPF tool is illustrated in the following steps.

(1) *Initialization*: After adding the necessary data layers (i.e., raster covariates layers and point response data) to ArcMap, the user starts the RSPF tool by clicking the RSPF button in the EAGLE Tools toolbox. Once open, the RSPF user interface shows three tabs, each containing drop-down boxes that users populate with appropriate map layers. These drop down boxes allow users to specify parameters including (a) Use and Availability layers, (b) the model region of interest, ROI, (c) the model spatial Resolution, and (d) model covariates.

To run the tool, a selection must be made in each of these boxes and an output folder must also be designated. If Use and/or Availability points fall outside the boundary of the selected ROI, the 'Clip Use, Availability, and Universe layers to this ROI:' checkbox should be checked. This limits model building to only those points that overlap the ROI. In addition to the option to provide an existing Availability layer, there are options for randomizing Availability points within a specified distance of a Use point and within a specified polygon layer. Detailed information is provided for each tab and drop-down box in a window on the right-side of the tool. Once all fields are filled in the user continues by clicking the submit button.

The RSPF tool then begins to create the input files necessary for R to build an RSPF model. Warnings are given as message boxes under certain circumstances, e.g., for a

Resolution layer containing grid cells that are not square, and provide an opportunity to exit the program. Additionally, processing is automatically stopped and the user is returned to the GUI forms when data are insufficient for processing (e.g., a covariate layer is not spatially referenced). In the event that this occurs, details are provided.

(2) *Define Data* Types: The next screen the will ask users to specify whether the data layers are continuous or categorical in nature. The default data type is continuous. Once all layers are correctly attributed the user clicks another submit button.

(3) *Prepare Data and Call R Script #*1: Model building begins by creating a Run folder, including the subfolders, CovariateGraphs, Parameters, Results, and Tables, in the specified Output Folder. When a Run folder already exists, the smallest available integer is added becoming, for example, folder Run1. Similarly, a second folder is created for temporary files and is named either TmpRSPF or TmpRSPF followed by an integer.

If point layers are to be clipped, the ROI layer is converted to a polygon layer and two new layers will be created, the intermediary ROIRecls, and ROIReclsPoly. A 'universe' point layer, RandUnivPts, is then created by randomly generating 10,000 points within the extent of the selected ROI. If a randomized Availability layer is to be created, it is done at this point. Randomization of Availability points within a specified distance of Use locations will result in a BufferUse# layer, where # is the specified distance. Both options for randomizing Availability points result in the layer RandAvailPts.

Coordinate pairs are added to each point in the Availability, Use, and Universe layers and are used to extract the covariate values at each location. If the 'Clip Use, Availability, and Universe layers to this ROI:' box is checked, these layers are clipped to the ROI. The tables for either the clipped or original layers are then exported as comma delimited files and stored within the Tables folder. A parameter file named RSPF_params_aks.txt (Parameters folder) is then generated. This parameter files contains the necessary folder information (i.e., paths to the datasets) and quantitative details about the raster covariates that are utilized by the first R script.

(4) *Define the Final Model and Call R Script #*2: The first R script produces information for each of the covariates selected on the Covariate Data tab in the first GUI. This resulting information is displayed in a third GUI where each covariate is contained within a unique tab. On each tab, users can select the fit order (first, second, or third) that best fits the data, or can choose to exclude the variable from further analysis. Several informative graphs (stored in the CovariateGraphs folder) are also provided on each tab to help the user determine whether variables are appropriate for the analysis or not. Another option on each tab (i.e., the "Swat Tool") allows users to apply the model to a different layer. Each of the Swap Tool drop-down boxes defaults to the layer on which the model was built. At this stage, users must also select a link function (i.e., logit, exponential, or loglog) for the model. Unlike the fit order, a single link function is applied to all variables in the model. Once all variables are selected the user clicks submit again, causing ArcGIS to create a second parameter file and then call the second R script.

(5) *Produce the Final Model Results*: The second R script creates several files in the Results folder including text files for the model equation, betas, and fit summary, as well as the rspf_ROC_semivariogram.jpg and rspf_rspf_resids.jpg visual diagnostic plots.

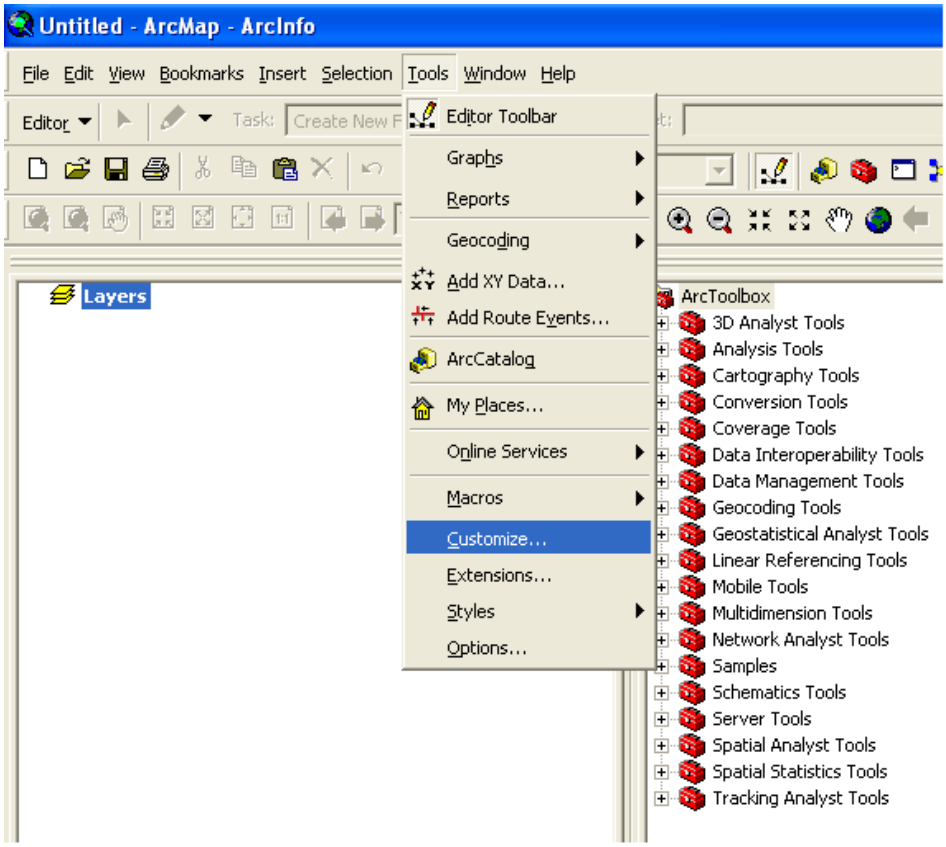
Finally, the MapAlgebra functionality native to the ArcGIS software package is used with the rspf_equation.txt file to create a response surface. The resulting response surface, in addition to the diagnostic graphs and tabular outputs produced by R, are the final results of the RSPF analysis and thereby represent the information from which inferences can be drawn.

Appendix 4: Installing the RSPF Tool as a Button

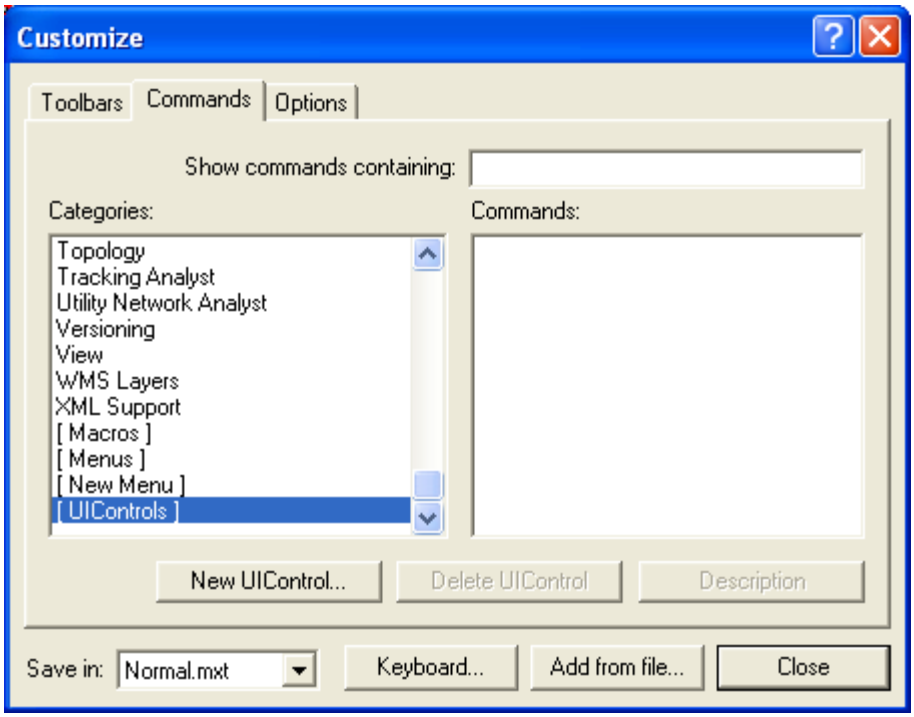
The RSPF tool is typically distributed within an ArcGIS project (.mxd) containing the necessary code for the Graphical User Interfaces (GUI) that constitute the tool. Alternatively, the RSPF tool can be installed as a clickable button that remains within a toolbar in the ArcMap environment. This document outlines steps that can be used to install the RSPF tool in ArcMap version 9.3. When you have completed these steps, the new button will appear in all new projects. Removal instructions are given at the end of the document. Note that the files necessary to install the button are provided with the other downloadable materials.

To install the RSPF tool:

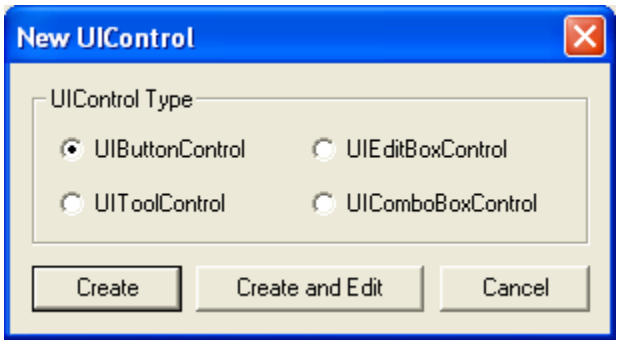
1) Under the Tools menu item, select Customize... Customize… is also available in the list that appears when right-clicking on a gray area in one of the menu bars.



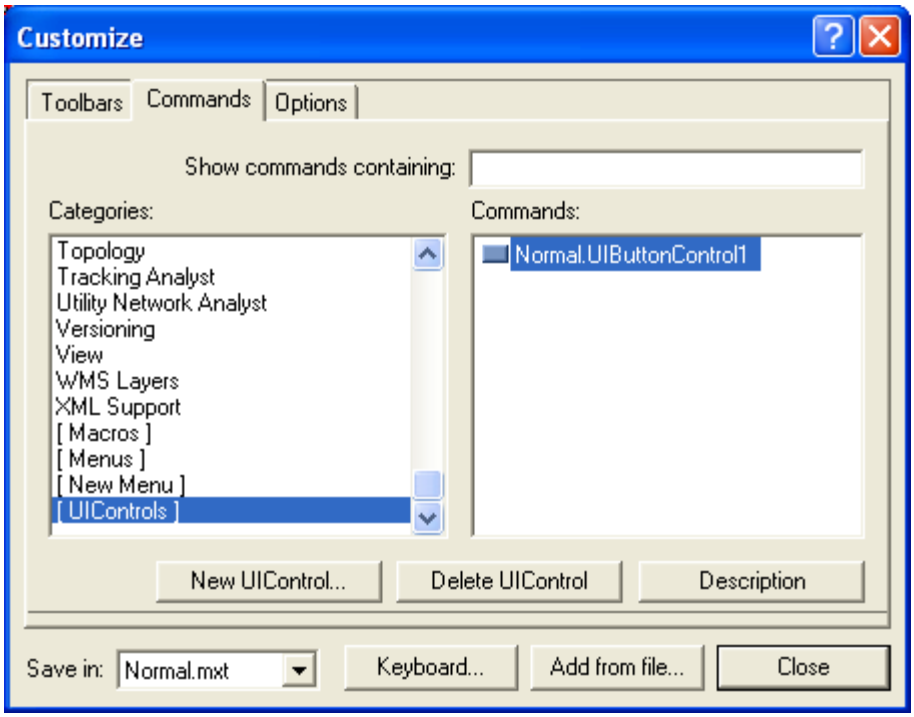
2) This will open a Customize window. Select the Commands tab and then [UIControls].



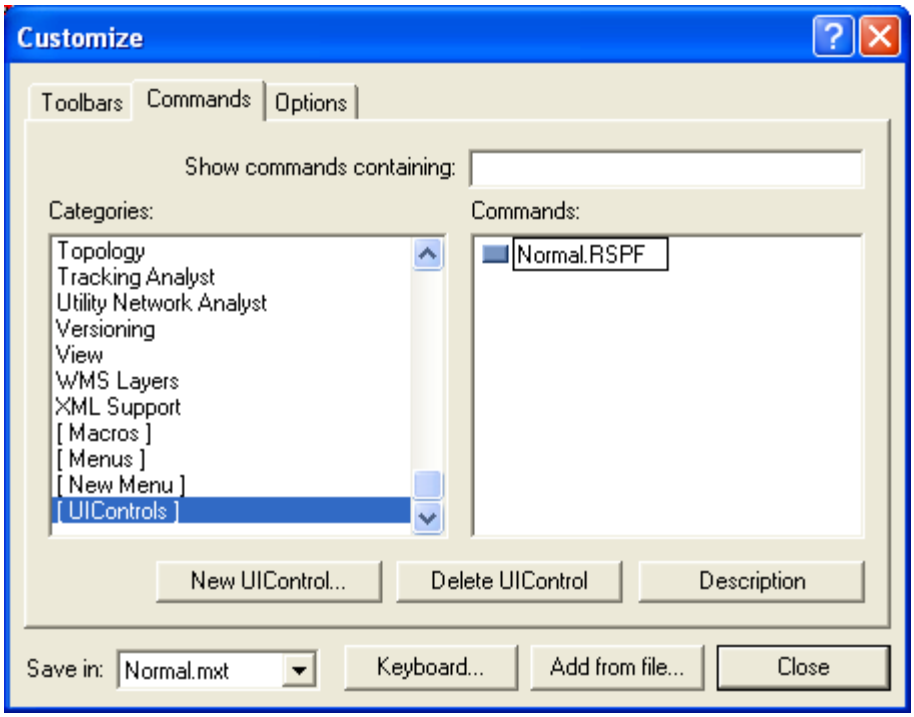
3) Click the New UIControl… button. In the New UIControl window that appears, select UIButtonControl and then the Create button.



The following window should appear. Notice the new button listed in the Commands: pane, Normal.UIButtonControl1.

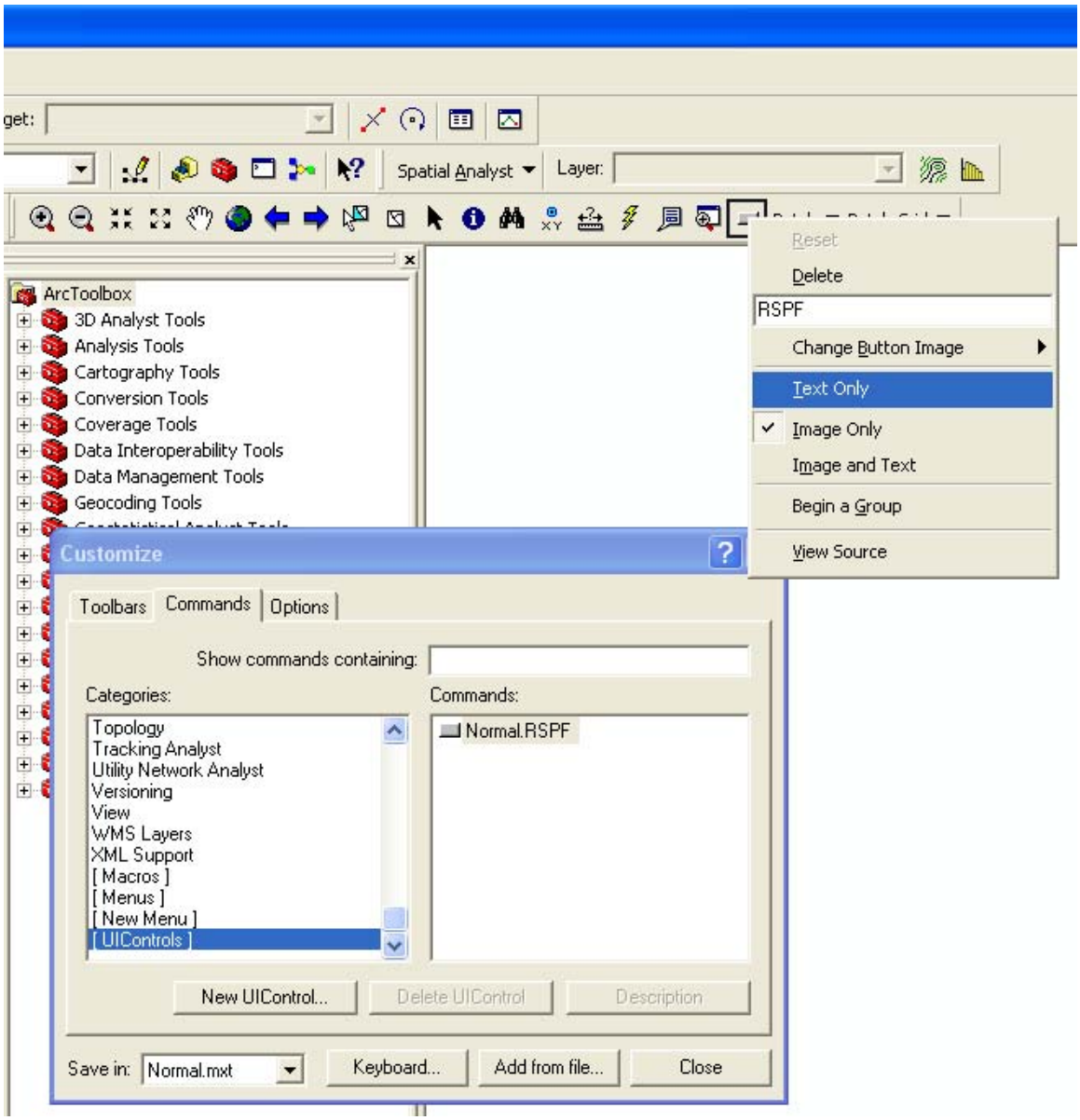


4) Click once on Normal.UIButtonControl1, wait briefly, and then modify the name of the button beyond the Normal. prefix. This will be the identifier for the button that you will be creating.

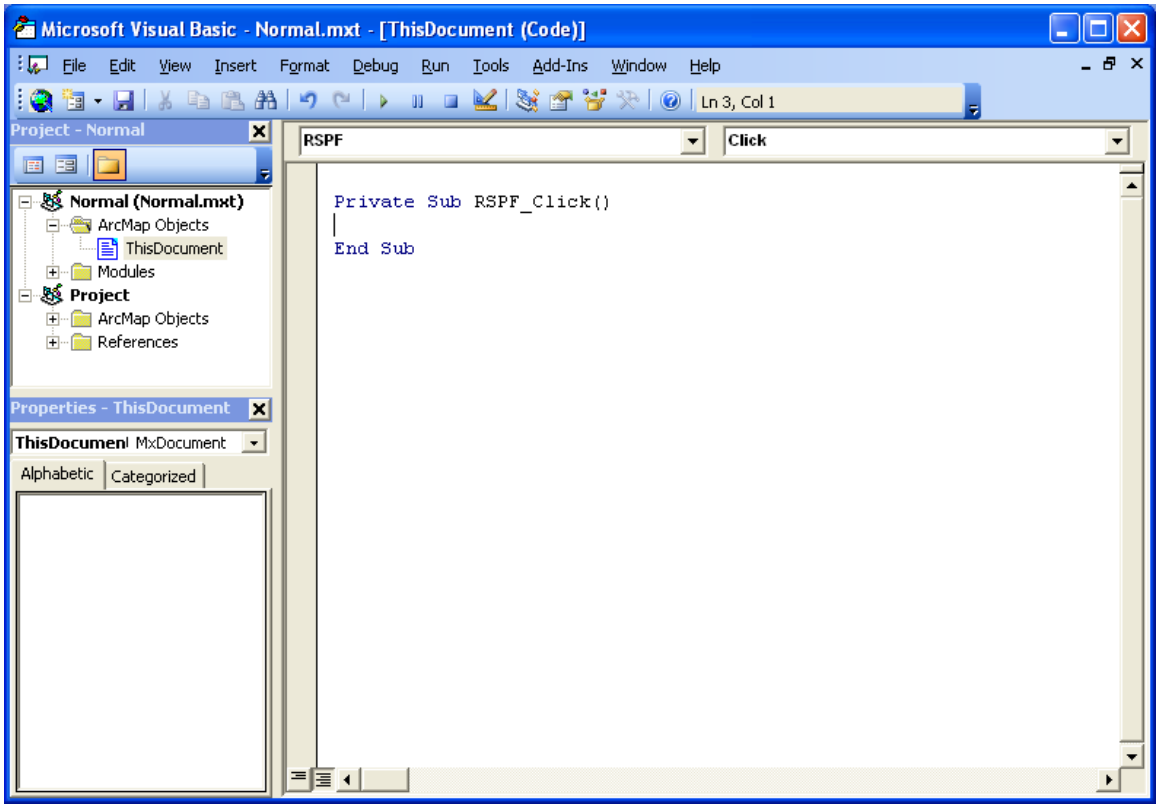


5) Click away from the name to exit the renaming mode. Then, click and drag the button to a menu bar.

6) Right-click the new button and select Text-only.



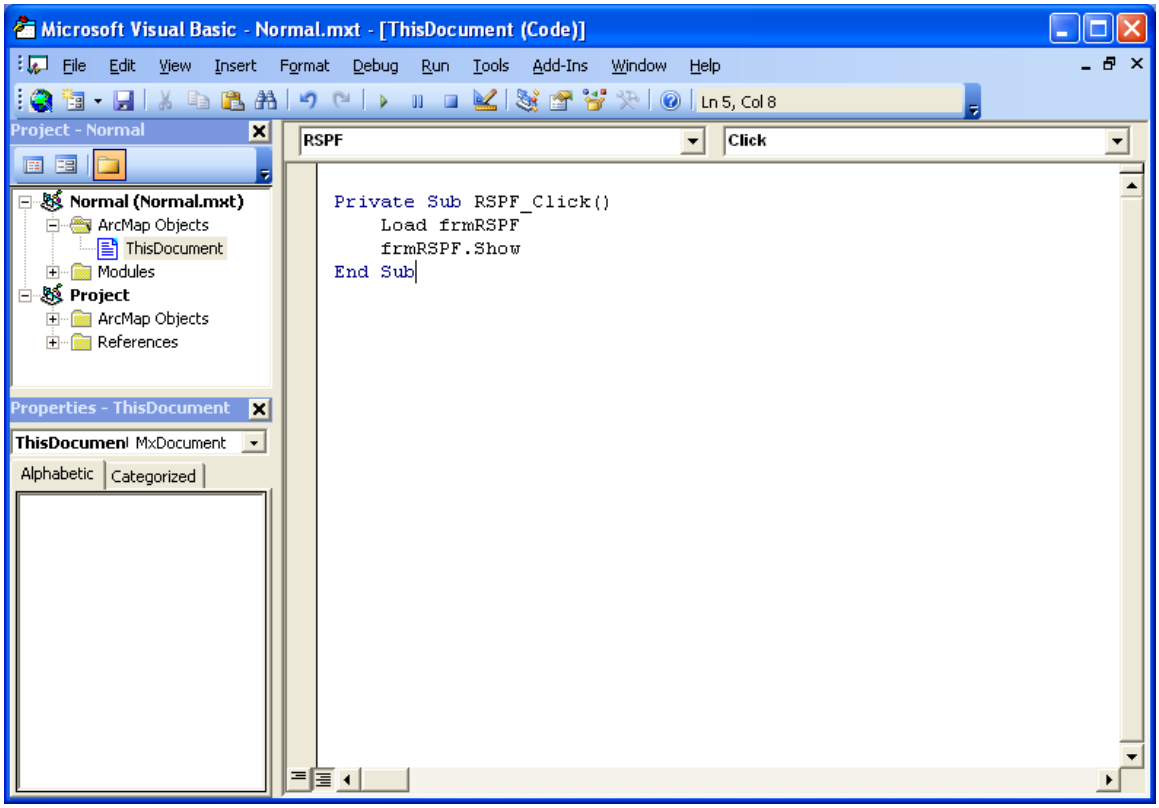
7) Back in the Customize window, double click the renamed button to launch Microsoft Visual Basic.



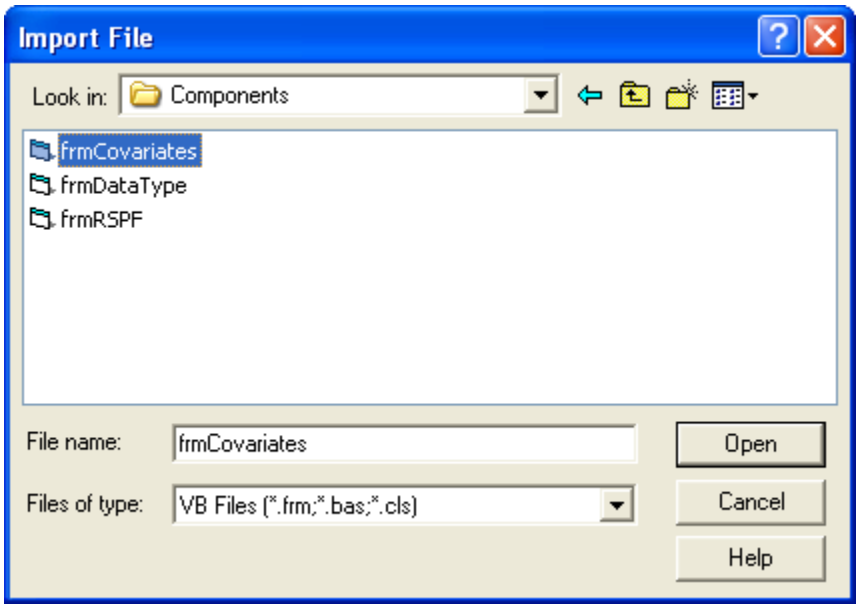
8) Add the following lines of code.

Load frmRSPF

frmRSPF.Show



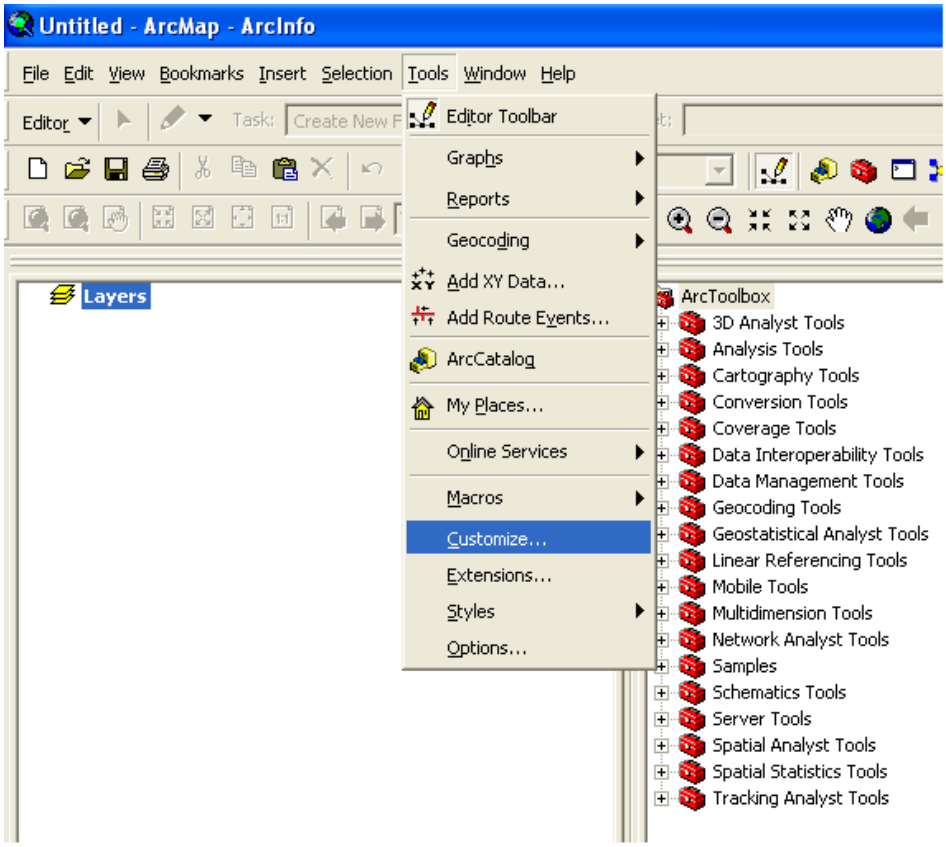
7) Under the File menu item, select Import File… and navigate to the directory where you unzipped the file. Repeat these steps three times and Open each of the forms, frmCovariates, frmDataType, and frmRSPF.



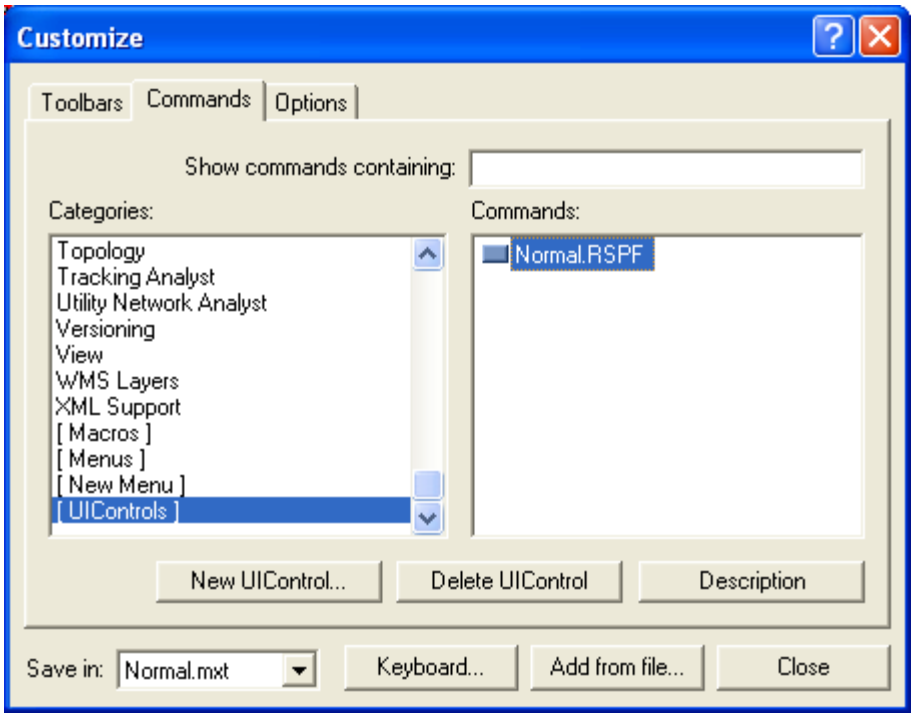
8) Under the File menu item, select Save Normal.mxt. You have now created and saved the new button to the master ArcMap project. Close the Microsoft Visual Basic window. Back in ArcMap, your new tool should appear as a button and be completely launchable.

To uninstall the RSPF tool:

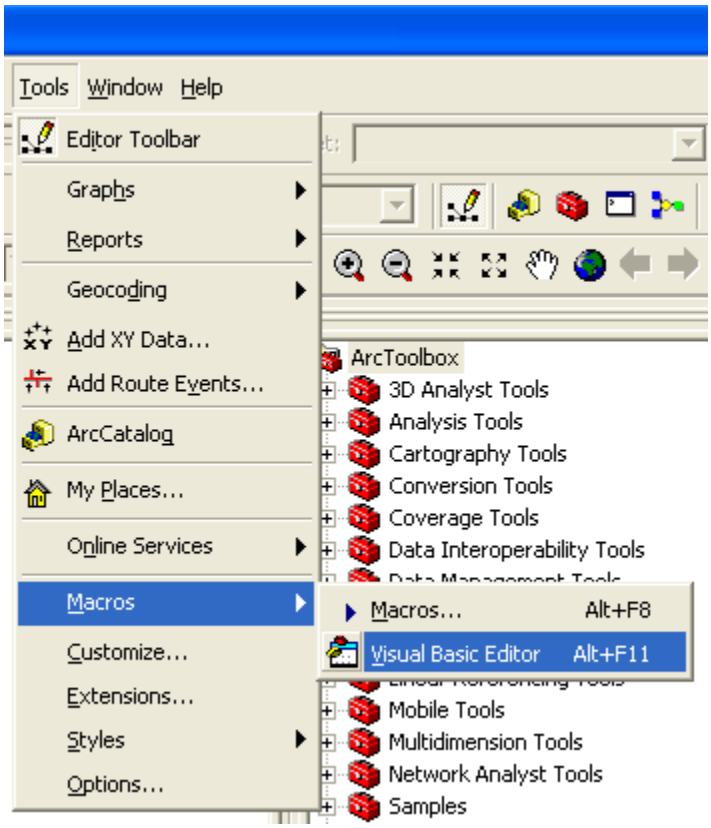
1) Under the Tools menu item, select Customize... Customize… is also available in the list that appears when right-clicking on a gray area in one of the menu bars.



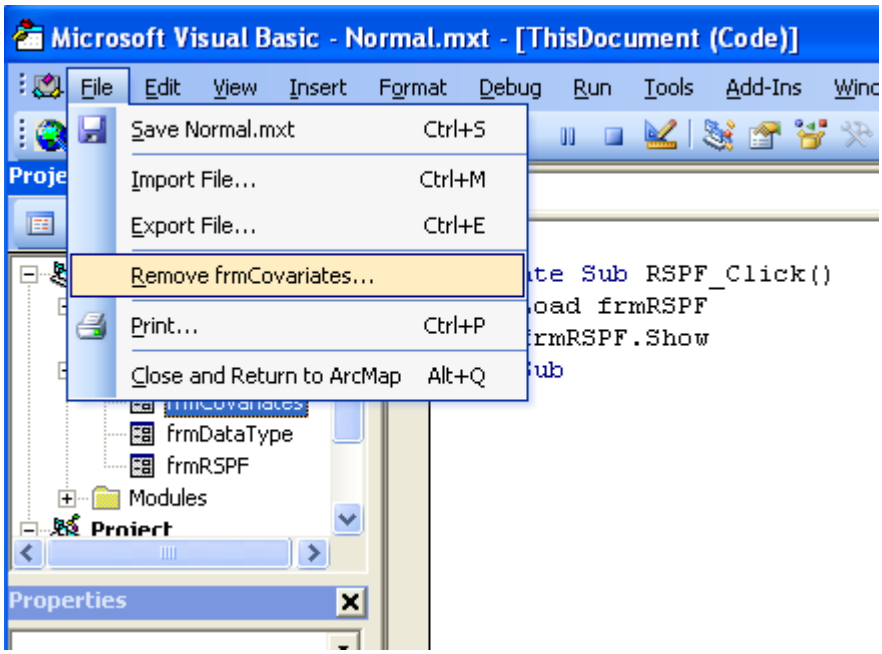
2) This will open a Customize window. Select the Commands tab and then [UIControls]. The RSPF button will appear in the Commands: window pane. Select it and then click the Delete UIControl button. Then click OK.



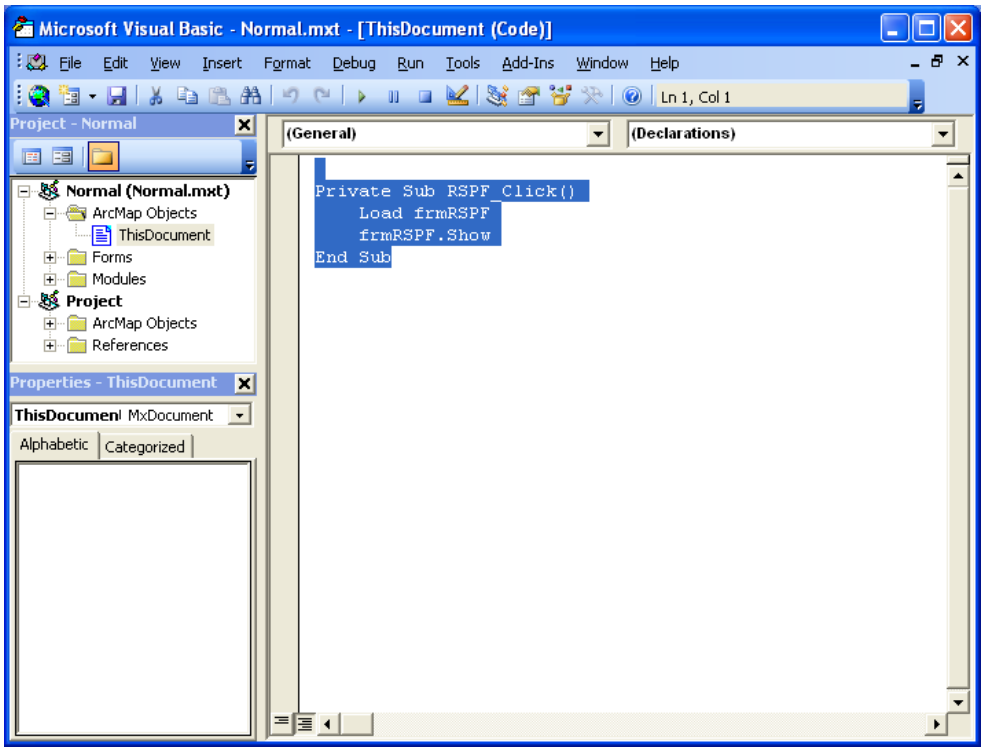
3) To remove the code associated with the button, return to Microsoft Visual Basic by clicking the Tools menu item, then Macros, and finally Visual Basic Editor.



4) Remove each of the three forms frmCovariates, frmDataType, and frmRSPF by selecing File, Remove frmCovariates…, etc.



5) Finally, delete the code in ThisDocument located under the ArcMap Objects folder.



Appendix 5: RSPF Analysis Key Questions

The goal of EAGLES is to augment existing decision support systems by providing tools that produce results that aid resource managers. The Resource Selection Probability Function (RSPF) Tool available in EAGLES provides a habitat selection model capable of producing results within a standardized and transparent framework. This appendix provides a rough guide for documentation we suggest accompany any report and/or publication in which results from RSPF are used. Note that these questions should be answerable *regardless* of the habitat selection model (e.g., MaxEnt, RSF, etc.) used to support management decisions.

1. What is the response sample size? How was this response obtained (i.e., via probabilitistic sampling of individuals, via appropriate temporal thinning of repeated measurements on the same individuals, etc.)? Are there inherent biases that might exist due simply to the collection of response data?

2. At what temporal scale (i.e., over what temporal period) were the responses collected?

3. What is the temporal extent of the analysis and are the response and covariate datasets synchronous in time? If not, how was the temporal domain for each covariate chosen (e.g., were certain temporal lags in covariate values included, and with what justification)?

4. What is the spatial extent of analysis and do the response data sample this area sufficiently to justify inference over the entire domain?

5. What is the spatial resolution of the analysis, how was this resolution arrived at (i.e., were covariates scaled up or down), and what ramifications will this resolution have on the results and their interpretation?

6. Are there any known and important covariates (i.e., raster dataset capturing a biophysical landscape characteristic of interest) that are missing from the analysis and what ramification(s) is this likely to have on the results?

7. How is availability space defined and what criteria were used to make this decision? If an iterative process was used to define availability space, what range of values was tested? How robust is model inference to changes in the availability space?

8. If RSFP results are extrapolated in time and/or space (i.e., used to make inference in different times or places from those in which the model was built) how is this decision justified? Do extrapolated spatial values represent covariate combinations actually observed in the extant data?

9. For each covariate what order fit was used and why?

10. What link function was used to produce the RSPF fit (i.e., the final modeled surface) and why?

11. Was spatial autocorrelation present in model residuals, and if so, what course of action were taken?